CHALMERS
UNIVERSITY OF TECHNOLOGY

# Reverse Transcriptase Characterization and Engineering for Genome Editing

Master's Thesis in Biotechnology

EMY SAETRE

DEPARTMENT OF LIFE SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's thesis 2026

# Reverse Transcriptase Characterization and Engineering for Genome Editing

EMY SAETRE


CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Life Sciences
Chalmers University of Technology
Gothenburg, Sweden 2026

Reverse Transcriptase Characterization and
Engineering for Genome Editing
EMY SAETRE



Supervisor: Aleksandar Zdravković Assoc. Prin. Sci., AstraZeneca
Examiner: Florian David Assoc. Prof., Department of Life Sciences, Chalmers University of Technology

Master's Thesis 2026
Department of Life Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Illustration of a reverse transcriptase.

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2026

Reverse Transcriptase Characterization and Engineering for Genome Editing
EMY SAETRE
Department of Life Sciences
Chalmers University of Technology

# Abstract

Templated editing, consisting of a Cas9 enzyme coupled to a reverse transcriptase (RT) directed by a programmable guide RNA to install edits in the genome, offers unprecedented potential for treating genetic diseases. Yet, translation to clinical applications is hindered by unpredictable editing efficiency and limited understanding of how RT properties influence outcomes. This thesis systematically characterized a panel of established and novel RTs, including engineered *Moloney Murine Leukemia Virus* (MMLV) derivatives, compact bacterial RTs, and AstraZeneca-proprietary candidates, through complementary cellular and in vitro assays to establish rational engineering principles. A HiBiT reporter system enabled high-throughput quantification of templated editing (TE) efficiency *in cellulo*, while biochemical assays were used to measure polymerization activity, DNA/RNA binding affinity, thermostability, and aggregation propensity. Results suggested that successful templated editing requires optimization of three critical properties: thermostability, protein solubility, and DNA/RNA substrate affinity. Two iterative rational engineering cycles based on this hypothesis highlighted a synergistic relationship between the three properties. Engineered variants that combine solubility, stability and substrate affinity enhancements achieved up to 2.4 times higher editing efficiency compared to the established state-of-the-art benchmarks. By systematically defining the biochemical and biophysical property thresholds required for TE efficiency, this work establishes an evidence-based design framework that transform RT development from empirical screening toward rational, predictive engineering, diminishing optimization timelines and accelerating therapeutic development.

Keywords: genome engineering, reverse transcriptase, protein engineering, rational design, templated editing, prime editing, ProteinMPNN, biochemistry

# Acknowledgements

Many people were involved in this project to different extents, and their expertise and support were invaluable for bringing this project together. First and foremost, I want to extend my most heartfelt thanks to my supervisor Aleksandar Zdravković who at times I believe was more excited about the results and progress than me. Without his guidance, unwavering support and general hype, I would not have made it anywhere. So thank you Aleks, for believing in me and giving it your all to support me even though you were extremely busy with your own things. Next, I want to thank Saša Svikovic for believing in this project, lending your wisdom to improve assays, and coming with brilliant ideas and feedback that enabled this project to grow and produce meaningful results. I also want to extend many heartfelt thanks to Euan Gordon and Lovisa Holmberg for their support with the purification of the proteins, and for teaching me their craft. I also want to thank Louis Dacquay for establishing the reporter assay I have been using throughout this thesis and designing the guides. To Joanna Szydzik who shared all of her tips and tricks for western blotting with me. To Anja Göder for being a fantastic friend, and helping out with her extensive expertise in automation and problem solving whenever I needed someone to discuss with. And to the rest of the people of the GE department for welcoming me to the team and always being so helpful. Lastly, I want to thank my co-writers and support team Tofu and Isa, for voicing their opinions on my writing with insistent key pressings, and giving up their time to proof-read right when it matters the most - while typing. Their contribution to the final product has been invaluable.

Emy Saetre, Gothenburg, April 2026

# List of Acronyms

Below is the list of acronyms that have been used throughout this thesis listed in alphabetical order:

| | |
|---|---|
| AAV | Adeno-Associated Virus |
| DMEM | Dulbecco's Modified Eagle Medium |
| DSB | Double Stranded Break |
| DTT | Ditiotreitol |
| Ec48 | *Escherichia coli* retron 48 |
| GAPDH | Glyceraldehyde-3-Phosphate Dehydrogenase |
| GOI | Gene of Interest |
| gRNA | guide RNA |
| HDR | Homology-Directed Repair |
| HEPES | 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid |
| Laf bench | Laminar Air Flow bench |
| LB | Lysogeny Broth |
| MIF1 | Macrophage Migration Inhibitory Factor 1 |
| MMLV | Moloney Murine Leukemia Virus |
| MMR | Mismatch Repair |
| mP | milli Polarization |
| MPNN | Message Passing Neural Networks |
| NHEJ | Non-homologous End Joining |
| nTPM | Normalized Transcripts per Million |
| PBS | Context dependent: Primer Binding Site (Templated editing) or Phosphate-Buffered Saline (Cell culture) |
| PE | Prime Editor |
| pegRNA | Prime Editing Guide RNA |
| PEn | Prime Editing Nuclease (Double-stranded DNA nuclease-based Templated Editing) |
| RLU | Relative Light Units |
| RT | Reverse Transcriptase |
| RTT | Reverse Transcription Template |
| TB | Terrific Broth |
| TCEP | Tris(2-carboxyethyl)phosphine |
| TE | Templated Editing/Templated Editor |
| Tf1 | Transposon of Fission yeast 1 |
| Tx100 | Triton-x100 |
| wt | wild type |

# Contents

# List of Figures

# List of Tables

# 1

# Introduction

In 2019, 200 million people around the world were estimated to live with a rare genetic disease [1]. Many common prevalent disorders such as cardiovascular disease, have also long been suggested to stem from genetic variations [2]. Since the first successful ectopic expression of a foreign gene in human cells in the early 1970s [3], successive generations of gene therapy technology have increased the efficiency, specificity, and safety of gene transfers. The discovery and development of RNA-guided nucleases, most prominently CRISPR/Cas9, in the mid-2010s made these gene editing reagents drastically simpler and cheaper to develop, fueling basic discovery and accelerating the development of therapies [4]. In principle, CRISPR/Cas9 paired with an exogenous DNA template, can perform endogenous gene editing, relying on the homology directed repair mechanism induced by double-stranded breaks (DSB) to accurately install intended edits at almost any site in the human genome [5].

Yet, for many therapeutic applications, conventional CRISPR strategies that rely on DSBs face persistent challenges, including unpredictable indel formation, low efficiency of homology-directed repair in clinically relevant cells, and cell cycle dependence [6][7][8]. These limitations have motivated a new generation of editors that minimize or avoid DSBs while improving precision and versatility. Among them, templated editing stands out for its ability to introduce any type of edit without relying on the DSB repair pathways [9][10].

Templated editing (TE) couples a programmable Cas9 enzyme to a reverse transcriptase (RT) and uses a specialized guide RNA that both specifies the genomic locus and encodes the desired edit [9]. After target recognition and nicking, the RT inserts the edit by extending the 3' end of the nicked genomic DNA using the gRNA's reverse transcription template (RTT). The edit will then be installed in the genome by the DNA mismatch repair system. This architecture has clear therapeutic appeal: it broadens edit scope, reduces imprecise edits due to DSB breaks, and has shown proof-of-concept in diverse cell types and loci. Nevertheless, gaps remain between laboratory performance and the demands of *in vivo*, clinical editing. For TE to advance toward the clinic, editors must be efficient, precise, programmable across diverse genomic contexts, and deliverable to target tissues. The large size of current editors challenges viral delivery. Methods such as splitting the editor and minimize the designs help, but often at a cost to activity [4]. Editing outcomes

vary markedly across sites and cell types, driven by gRNA structure, chromatin environment, and DNA repair dynamics.

A central, underexplored determinant of TE performance is the RT component. Most early systems adopted *Moloney Murine Leukemia Virus* (MMLV) RT variants engineered for higher stability and binding [9], and later work introduced evolved derivatives of MMLV RT and also explored smaller RTs from bacteria (Ec48) and yeast (Tf1) [11]. These efforts demonstrated that changing the RT can shift efficiency, increase the possible sequence length that can be inserted using TE, and compatibility with different RTT structures. The RT's intrinsic properties likely set the ceiling for how robustly RTTs are copied and installed. However, despite these advances, our mechanistic understanding of how RTs affect TE outcomes in cells remains limited. A systematic dissection of RT attributes could illuminate levers for improving TE and enable effective rational engineering.

## 1.1 Aim

This thesis aims to bridge the gap between reverse transcriptase (RT) functional properties and templated editing (TE) outcomes to move TE from empirical optimization toward informed rational design. The resulting insights would enable rational engineering of RTs that are more efficient, robust across loci, and that can insert longer sequences into the genome, features that are directly relevant for therapeutic delivery and efficacy.

### 1.1.1 Objectives

To achieve the aim, a panel of established and novel RTs, including engineered *Moloney Murine Leukemia Virus* (MMLV) derivatives, compact bacterial RTs, and AstraZeneca-proprietary candidates was benchmarked across complementary cell- and *in vitro* assays to connect cellular editing outcomes to quantifiable RT properties. The attributes that were investigated through cellular assays involved RT effect of TE activity under different conditions, while RT biophysical and mechanistic properties such as stability, polymerization activity, processivity, and substrate binding were benchmarked with biochemical *in vitro* assays. Finally, RTs were rationally engineered to improve thermostability, substrate affinity and solubility utilizing the obtained data to make informed decisions. The following objectives were established to complement the aim:

- Establish a reporter system for accurate, fast, and high throughput readout of TE activity in the cell line HEK293T and benchmark TE efficiency, insertion length, and genomic site dependence.
- Establish *in vitro* assays to benchmark RTs' polymerization activity, substrate binding, thermostability and aggregation.
- Rationally engineer RTs through site-directed mutagenesis and AI pipelines, informed by cellular and *in vitro* assays to improve TE activity.

# 2

# Theory

This chapter will cover the theory of the project, aiming at providing the right level of understanding to follow later chapters. Genome editing and its history, including CRISPR/Cas9, will be introduced to explain the novelty of templated editing. Reverse transcriptase structure, mechanism, and families will be covered to explain the RTs used in this project. Lastly, this chapter will explain assay designs and methods.

## 2.1 The History and Practice of Genome Editing

Gene editing is a technology that precisely modifies the genome sequence to induce insertions, deletions, or base substitutions in the genome [12]. The discovery and repurposing of nucleases was the beginning of programmable gene editing, starting with the development of zinc-finger nucleases (ZFNs), followed by transcription activator-like effector nucleases (TALENs) and meganucleases [4]. Both ZFNs and TALENs use proteins to identify and target DNA strands. While these nuclease-only-based techniques evoke low immune response, the precision needed for the proteins to recognize the desired DNA sequence makes the ZFNs and TALENs difficult and time-consuming to design, and often results in inefficient editing [13]. It was not until the discovery of the programmable RNA-guided CRISPR/Cas9 system that precise genome editing and gene therapy became widely feasible, revolutionizing biomedical research and accelerating the development of novel treatments due to its low cost, ease of use, and high efficiency.

### 2.1.1 CRISPR/Cas9

CRISPR/Cas9 is a highly effective genome editing tool that was discovered in 2012 by scientists Emmanuelle Charpentier and Jennifer A. Doudna [14]. The CRISPR/-Cas (Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR-associated) system evolved naturally in bacteria and archaea as a defense mechanism against phage infection and plasmid transfer. The defense system provides the bacteria with immunity through the integration of foreign genetic material into the chromosomes, allowing targeted silencing and degradation upon repeated infection [15].

The defense system is composed of cas genes organized in operons, and a CRISPR array consisting of unique genome-targeting sequences called spacers, interspersed with identical repeats [14]. The system relies on small RNAs for detection and silencing of foreign nucleic acids. The mechanism for immunity is initiated by integrating short fragments of foreign genetic material (protospacers) into the chromosome at the proximal end of the CRISPR array. When the sequence has been integrated into the genome, it is called a spacer sequence. Upon transcription of these spacer sequences along with the repeat region, CRISPR RNA (crRNA) is created. The crRNA is capable of base pairing with the complementary protospacer sequences of invading viral or plasmid targets. Elimination of the foreign nucleic acids will be achieved through means of Cas enzymes that function in complex with the crRNAs.

There are two classes of CRISPR/Cas systems separated based on the number of effector proteins required to identify and cleave target nucleotides [16]. The first class uses a complex of multiple Cas proteins, while class 2 only requires one Cas protein, making it more relevant for gene editing purposes due to its simplicity. There are six sub-types of CRISPR/Cas systems, I-IV. Type I, III, and IV belongs to class 1, while type II, V, and VI belongs to class 2. Type II systems are the most widely used for genome editing because of their high efficiency and convenience [16]. The type II CRISPR/Cas9 system is comprised of three components: the crRNA, a trans-activating crRNA (tracrRNA), and a Cas9 protein. These three components enables specific targeting of sequences through Cas9 mediated cleavage. For this type of cleavage, a guide is needed for the Cas9, along with something that attaches the guide to the Cas9. These functions are fulfilled by the crRNA and a tracrRNA respectively. The tracrRNA will bind to the Cas9 and pair to the repeat region of the crRNA allowing it to attach to the Cas9 to provide specific sequence targeting. The crRNA and tracrRNA was successfully fused to produce a chimeric single guide RNA (sgRNA) early in the development of the CRISPR/Cas9 editor [14]. Site-specific recognition occurs at locations determined by both base-pairing complementarity between the sgRNA and the target protospacer DNA, and a short PAM motif (protospacer adjacent motif) that sits next to the complimentary protospacer on the target DNA. Upon binding the target sequence, Cas9 will cleave the DNA three bases upstream of the PAM sequence by introducing a double-stranded break (DSB). If at this point, a DNA donor sequence with overhangs matching the break site is available, the homology-directed repair (HDR) mechanism can repair the DSB by introducing a precise insertion at the site using the donor sequence as a template [12]. Alternatively, the non-homologous end joining (NHEJ) will ligate the DSB. The NHEJ pathway has been particularly useful in genetic knock-out experiments and functional genomic screens, while the HDR pathway is especially appealing to exploit for clinical purposes as the mechanism allows error-free repair and insertions [7].

Until recently, the CRISPR/Cas9 system has been considered the best choice for genome engineering, however, the technique faces some challenges for applications in therapy due to the risk of off-target effects and indels [6]. The DSBs induced by CRISPR/Cas9 can also lead to unintended deletions or rearrangements [17][18].

Although HDR pathways can facilitate a desired edit, the efficiency of the pathway is limited, restricting its utility in precise gene editing for clinical intervention [7]. NHEJ occurs more frequently than HDR in most cell types but often involves insertions and deletions of base pairs at the cut site. The dependence on HDR for precise editing also renders CRISPR-mediated editing dependent on the cell cycle as the cell cycle stage plays a key role in determining available DNA-damage repair pathways [8]. HDR events are generally restricted to dividing cells in the late S and G2 phases, whereas NHEJ predominates the G1, S, and G2 phases. The low effciency of HDR in most therapeutically relevant cell types makes it difficult to correct or install precise edits for clinical applications [9].

### 2.1.2 Next Generation Genome Editors

Engineering of the Cas9 and optimizations of the sgRNA have improved precision of the technique. Recent advancements that aim to circumvent the DSBs includes base editing and templated editing (TE). In base editing a catalytically inactive dead Cas9 (dCas9), or Cas9 nickase (nCas9) for higher editing efficiency, is conjugated to deaminase, which can catalyze the conversion of nucleotides via deamination [19]. Base editors provide a means to edit single nucleotides without the toxicity induced by DSBs, however, the applications are limited as base editors are restricted to substitutions of C to T or A to G depending on the type. More recently, TE has been developed as a strategy to edit the genome to insert a desired stretch of edits without inducing DSBs [9]. This technology combines fusion of nCas9 with a reverse transcriptase (RT) and a guide RNA (gRNA), which contains a sgRNA sequence, and an RNA template encoding the desired edit. In principle, TE mediates targeted insertions, deletions, all possible base-to-base conversions without requiring DSBs or donor DNA templates.

## 2.2 Templated Editing

Templated editing (TE) enables a range of different types of edits into genomes with minimal indels and off-target effects [9]. Since the first appearance of TE systems, coined prime editors (PE) by David Liu's research group in 2019 [9], TE has been adopted for a diverse range of applications, for example, genetic modeling and therapeutic genome editing. This section will cover the TE mechanism for gene editing, the evolution of TE from 2019 until today, and the limitations of TE for therapeutic applications.

### 2.2.1 Mechanism of Editing

TEs usually consist of a Cas9 nickase (nCas9), such as a SpCas9, with one of its endonuclease domains inactivated, conventionally fused to a reverse transcriptase (RT) through a flexible linker (*figure 2.1*). A guide RNA (gRNA), the usual structure that will be described here is often called pegRNA (prime editing guide RNA), functions both as a guide and as a template to introduce the edits. From the 5' to

the 3' end, the gRNA consists of a spacer, a scaffold, a reverse transcription template (RTT), and a primer binding site (PBS).

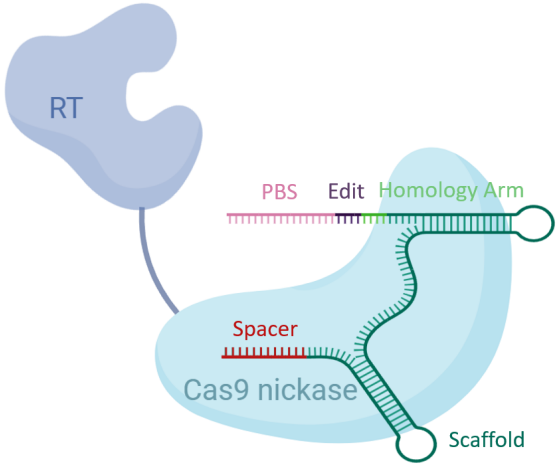



**Figure 2.1:** Templated editor (created in biorender.com).

To introduce edits, the Cas9 domain binds the gRNA's scaffold, whereafter the spacer sequence will direct the TE to the target genomic site that is defined by the combination of spacer complementarity, and having the correct downstream protospacer adjacent motif (PAM) sequence. The TE complex binds to the target DNA and Cas9 nicks the PAM-containing strand. The resulting 3' end hybridizes to the gRNA's PBS, then primes reverse transcription of new DNA containing the desired edit using the RTT of the gRNA.

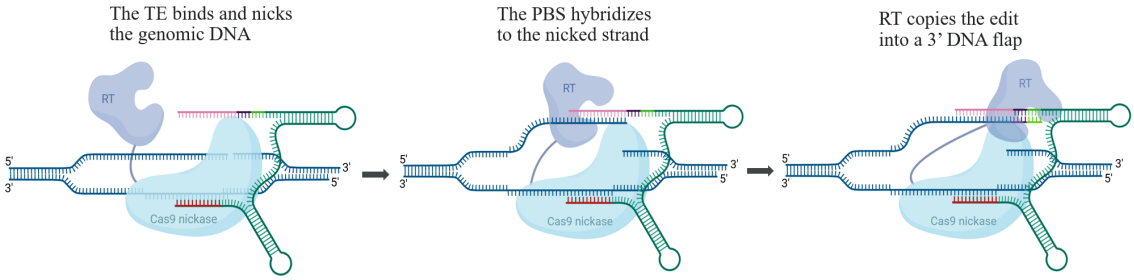



**Figure 2.2:** Templated editing conceptual model (created in biorender.com).

In addition to the desired edit, the RTT also contains a downstream homology arm to allow the edited strand to hybridize to the unedited complementary strand, designed to engage the cellular DNA mismatch repair system which is believed to facilitate incorporation of precise edits. Hybridization of the homology arm causes flap equilibration of the edited 3' flap and the unedited 5' flap [9]. The cell favors degradation of 5' flaps compared to 3' flaps, and subsequently, the unedited 5' flap will be degraded by native endonucleases, such as Flap Endonuclease 1 (FEN1) [10]. The resulting intermediate contains a mismatch and single strand break on

the edited strand. Due to the mismatch, enzymes from the cell's DNA mismatch repair (MMR) system will be recruited to the site to cut out one of the strands. There is a higher probability of cleaving the edited strand due to the nick already introduced by nCas9, which means removal of the edit and DNA repair according to the unedited strand. However, if DNA ligation of the nick occurs before the MMR system is initiated, the strand removal will depend on which strand contains pre-existing single-strand breaks generated independently from PE. Introducing a distal nick on the complementary strand will encourage the replacement of the unedited strand and thus promote installation of the edited sequence [9].

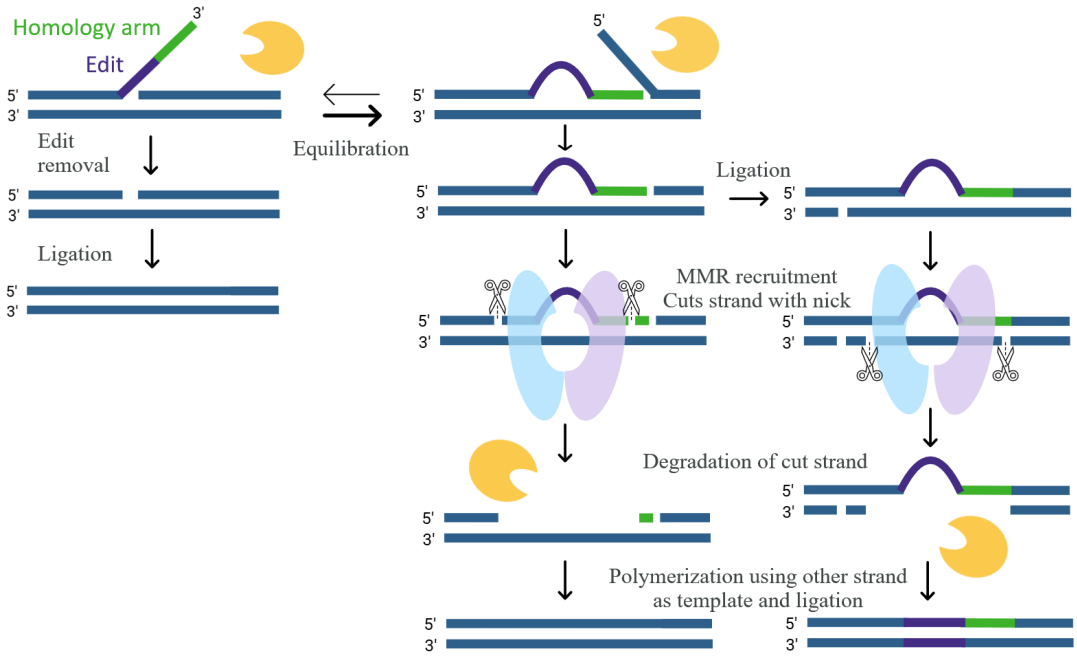



**Figure 2.3:** Schematic of edit incorporation with the MMR system (created in biorender.com).

### 2.2.2 Evolution of the Templated Editor

*Figure 2.4* features the major developments and milestones in the field of TE, relevant to the topic of this master's thesis project. Most variants were developed by David R. Liu's team, the creator of the PE system [9], however, there are other groups that have also made major contributions to the field by optimizing TEs and studying the editing mechanism. These include, for example, engineered gRNAs [20][21][22][23], compact TEs [24], untethered TEs [25], nuclease TEs [26][27][28], and overcoming PAM-sequence limitations [29][30].

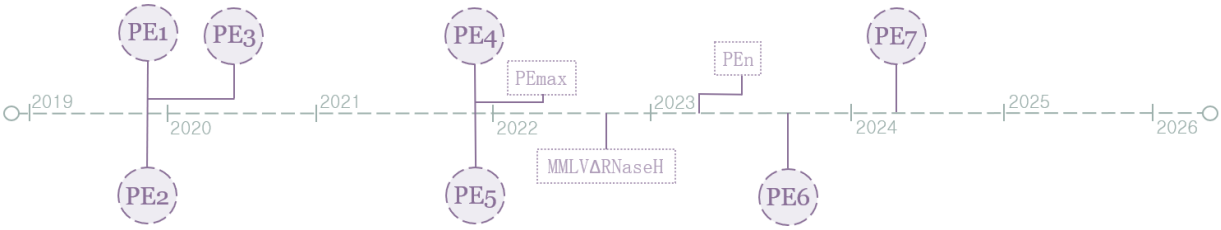



**Figure 2.4:** Timeline of the development of prime editors from 2019 until today.

PE1 was the first TE to be developed and it was designed as a proof of concept of the technique in cells [9]. The early editor consisted of a wild type MMLV RT fused, with a flexible linker, to the C-terminus of Cas9 H840A nickase. PE2 was created immediately after PE1 to improve the limited editing efficiency [9]. PE2 was developed by optimizing the MMLV RT component of the editor which successfully improved editing without increasing the number of unintended edits. Compared to PE1, *in vitro* editing experiments in human and mouse cells showed that PE2 has an increased range of editable genomic locations and is compatible with shorter PBS sequences. The third PE was created in conjunction with the second editor in an attempt to improve editing efficiency further through manipulation of gene repair outcomes to favor incorporation of the edit into the genome. In PE3, a second gRNA is introduced to direct the Cas9 to make a nick in the non-edited strand without causing DSBs. Identifying this nick in the non-edited strand, the MMR may more frequently replace the non-edited strand using the edited strand as a template, thus promoting installation of the edited sequence in both strands.

When developing PE3, the belief was that cellular MMR action would promote the installation of templated edits when nicking the non-edited strand [9]. However, further research into the mechanism of MMR and factors that impact TE efficiency revealed that MMR actually counteracts the efficiency of prime editing and promotes indel formation by the introduction of additional nicks in the edited strand along with degradation of the edited flap [31][32]. In response to this discovery, PE4 and PE5 aimed to manipulate the MMR pathways to improve genome editing outcomes [31]. The strategy used in both systems was the transient expression of a dominant negative MMR protein that acts by forming nonfunctional MMR complexes and disables the functional MMR proteins [33]. The reduced fidelity of the MMR mechanism could delay DNA repair or increase the likelihood of strand ligation, thus removing bias in the repair of the edited product.

This far, the RT component of the TE had not been optimized for human cells which caused some cell lines to exhibit lower expression of the editor, leading to reduced editing efficiency [31]. PEmax was developed to address this issue by an architectural improvement of the PE adding nuclear localization signals, an improved linker, and a codon optimized sequence for expression in human cell lines.

In an attempt to make TEs smaller to facilitate cellular delivery, many different

approaches were tested. Two of these strategies that were highly successful was to truncate the MMLV RT in PE2 to create the minimized MMLVΔRNaseH, which reduced the TE size to allow dual adeno-associated virus (AAV) delivery to cells, while still maintaining TE activity [24]. The other strategy split the RT and Cas9 domain of the editor into two separate plasmids to create a split TE system where the two proteins were not tethered [25]. Aside from enabling dual AAV delivery of the editor, the split system offers increased modularity. By splitting the proteins into different plasmids, it is possible to exchange any of them for others containing a variant protein. Exploring the split system and different minimized TEs, it was also discovered that the RT and Cas9 domain might possibly work in *trans*, instead of in *cis* as has been the working theory this far. If the two proteins work in *trans*, it would mean that tethering of the two components are futile, as another RT not tethered to the target site might provide the RTT extension. Comparison of the systems suggested that editing efficiency was comparable with the tethered and untethered PE2 system [25].

The discovery that the MMR system suppressed TE efficiency led to a desire to create a TE independent of the MMR. This was achieved by the creation of nuclease-dependent prime editors (PEn). The PEn editor connected a nuclease Cas9 to an RT and thereby caused DSBs in the genome when editing. Initial tests found that PEn was more efficient at DNA editing than nicking TE [26][27]. The downside is that indels and undesired edits are more common using this method. The method of repair of the DSB is still unknown at this point, though there has been observation of non-homologous end joining (NHEJ)-mediated integrations of gRNA RTTs during PEn editing, which has led to the development of a new type of guideRNA - the springRNA. Compared to the conventional gRNA structure described in *section 2.2.1*, sometimes referred to as pegRNA, springRNA does not require a homology sequence in the RTT as insertion is mediated through precise NHEJ [28].

To further improve editing efficiency of TE, David Lui's lab developed the sixth generation of PEs and introduced two new RTs and a novel approach to engineering through phage-assisted evolution which resulted in PE6a-g [11]. The work constituted the first large-scale screen of unknown RTs for TE activity, identifying 20 new RTs compatible with TE [11].

### 2.2.3 Limitations of Templated Editing

Despite substantial advancements, TE still faces challenges limiting its therapeutic potential. One of the major obstacles that makes TE less attractive for therapeutic applications compared to other genome editors, is the complex gRNA design that requires tailoring and optimization for the editing site and the Cas9 used [34]. TE also faces efficiency and specificity issues across diverse cell types and organisms with editing levels that vary significantly depending on the target site and the cellular context [33].

Another challenge that many research groups are actively addressing, but still re-

mains to be resolved, is the TEs' large size which limits the options for cellular delivery. Current gene delivery systems, such as AAVs, are limited by their small cargo capacity. The split TE systems, as well as more compact editors with minimized RT and Cas9 domains, aimed at solving this issue, however, optimization is still necessary to improve editing efficiency and reduce indels [35][33]. Nonviral delivery strategies such as nanoparticles or lipid-based delivery, offer an alternative but are still in the experimental phase, requiring further optimization for tissue specificity and efficient cargo protection [33].

## 2.3 Reverse Transcriptases

Reverse transcriptases (RTs) are DNA polymerases that can use RNA as a template for DNA synthesis, meaning that they catalyze the reverse of transcription [36]. For a long time, the central dogma of life was described as the unidirectional flow of information from DNA to RNA and finally to proteins. The discovery of RTs in the 1970 challenged this concept of unidirectionality [37]. The discovery of the RT, and subsequent research, revealed that reverse transcription not only plays a major role in the replication of retroviral genomes, but is also encoded in genetic elements of prokaryotes and eukaryotes [36]. In fact, RTs are involved in the conservation of telomeric ends on chromosomes [38], in the mitochondrial plasmid replication [39], and in processes of gene transposition [40].

The potential of reverse transcriptases in biotechnological and clinical applications was recognized early after it's discovery [36]. In combination with polymerase chain reaction (PCR), reverse transcription PCR (RT-PCR), RTs offer a powerful tool with nearly an unlimited range of applications such as detection of expression levels of particular genes, to generate cDNA libraries, to identify RNA levels in patient samples for clinical diagnostic or in newly developed next-generation RNA sequencing technologies [36]. During the Corona virus pandemic, RT-PCR became the gold standard for diagnosing SARS-CoV-2, as RT-PCR can be used to detect very small amounts of viral RNA in patient samples. Currently, most of the protocols aimed to analyze RNA transcripts involve the use of recombinant variants of murine leukemia virus (MMLV) RT, although avian myeloblastosis virus (AMV) and human immunodeficiency virus type 1 (HIV-1) RTs, among some others, have also been extensively characterized [41].

Recent applications of RTs include employment in sequencing and genome editing by careful engineering of appropriate properties. Classical retroviral RTs have been improved by protein engineering to develop next-generation sequencing techniques, RNA-sequencing, that are now being applied to the study of transcriptomics [41]. Progress in next-generation sequencing technologies has paved the way for single-cell transcriptomics, leading to a deeper knowledge of cellular heterogeneity. To maximize efficiency, these RT enzymes have been designed with increased catalytic efficiency at high temperatures, ability to synthesize long cDNA products, and high sensitivity [41]. In recent years, engineered RTs fused to a CRISPR/Cas9 nickase have shown great potential as tools to manipulate eukaryotic genomes in the

previously mentioned templated editing technique.

### 2.3.1 Structure and Domains

The typical RT consists of five domains: fingers, palm, thumb, connection, and RNase H domains (*figure 2.5A*) [42]. The structure exhibits a clamp-like shape for accommodating nucleic acid binding and shows some resemblance to a right hand, hence the names. RTs show high functional and structural similarity, although low sequence identity [43]. RTs usually combine two enzymatic activities located in two separate protein domains: a DNA polymerase activity that uses RNA or DNA as a template, and an RNase H activity responsible for the degradation of RNA templates when DNA/RNA hybrids are formed [36].

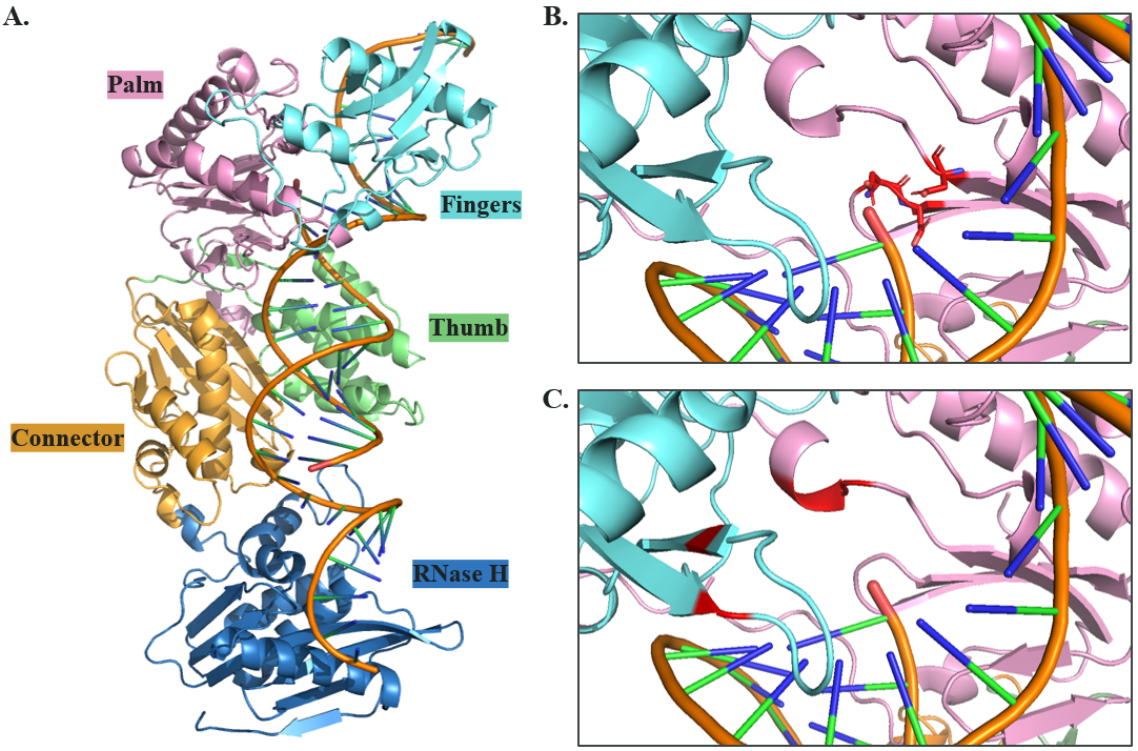



**Figure 2.5:** **A.** The domains of MMLV RT [42], (structure predicted with Boltz-2). **B.** The active site of MMLV RT (red). **C.** The dNTP binding residues of MMLV RT (red).

Fingers and palm are the most conserved domains, they contain the polymerase active site where elongation occurs, and are tightly folded and structurally stable to provide a cleft for substrate binding [42]. During reverse transcription, the finger domain will open and close with a hinge-like motion to incorporate nucleotides [44]. The fingers domain has been proposed to provide an intermediate binding site for the template in between phosphonucleotidyl transfer reactions [42]. The active site is comprised of three conserved aspartate residues and is located on a structurally conserved epitope in the palm subdomain consisting of three antiparallel β-strands

[45] (*figure 2.5B*). In addition, a conserved loop motif is thought to contribute to the proper positioning of the incoming dNTP with the substrate, while two residues in the finger domain help coordinate the triphosphate of the dNTP (*figure 2.5C*).

The thumb domain consists of two principal helices and is believed to play an important role in substrate binding and processivity, and has been suggested to provide a "minor groove binding track" important for translocation of the substrate during polymerization [43][45].

The RNase H domain is linked to the thumb domain through the connection domain which provides conformational flexibility. The connection domain's conformation is believed to significantly affect the interaction between nucleic acid and the finger, palm, and thumb domain along with also being critical for the hydrolysis of the RNA during polymerization [45]. The RNase H domain has polymerization-independent activity that will cleave DNA/RNA hybrids. In viruses this activity is combined with strand transfer events to degrade the RNA template after synthesis of the (+) strand to also synthesize the (-) strand to obtain double stranded cDNA [41]. The RNase H activity will also facilitate reverse transcription of, for example, hairpin structures, but can also cause unwanted degradation of the template [46].

Several conserved motifs have been found in RTs, including YXDD, TVLD, and LPQG [42]. YXDD is believed to be important for enzyme function and contains two of the three aspartates that constitutes the enzyme catalytic site. Mutations of the residue in the X position has been shown to drastically affect polymerase activity. Another conserved motif in some RTs is the DED motif which constitutes the RNase H active site [45].

### 2.3.2 Polymerization Mechanism and Influencing Factors

The mechanism of RT binding to a substrate includes multiple steps involving covalent bonds and conformational changes. The RT has two well characterized states: the open-finger conformation and the closed-finger conformation [44]. The open conformation is the native unbound state associated with a retracted fingers-, and opened thumb domain [47], leaving an open cavity that allows for binding of nucleotide substrates.

The open conformation allows rapid binding of dNTPs from solution, but only the complementary dNTP that can base pair to the template will have the right fit to allow the proper conformational change into the closed-finger formation [48]. An incoming dNTP unable to base pair with the template bases will sterically hinder the closed-finger formation and thus dissociate at an increased rate. This induced-fit mechanism ensures the correct nucleotide incorporation and accounts for the fidelity of RT polymerization. When the correct nucleotide enters the RT, the fingers subdomain moves by a combined hinge-like motion, closing around the incoming nucleotide to assemble the polymerase active site for nucleotide incorporation [44]. For this conformational change to take place, two $Mg^{2+}$ ions are also required in the

active site of the RT [49][44]. Other ions such as $Na^+$, $Mn^{2+}$, and $Ca^{2+}$ can also enter the active site, and might affect the rate and fidelity of the polymerization reaction [50][49].

While in the closed conformation, one $Mg^{2+}$ ion will be in contact with the three phosphates of the bound nucleotide, while the other is situated between the $\alpha$-phosphate of the nucleotide and the 3'-end of the substrate primer [44]. The polymerization process cycles between open and closed conformation during nucleotide addition. To catalyze the nucleotidyl transfer reaction of the dNTP, the $Mg^{2+}$ ion situated next to the 3'-end of the substrate primer will activate the OH group for attack of the $\alpha$-phosphate of the dNTP, while the other metal ion stabilizes the displaced pyrophosphate moiety. This will widen the minor groove of the substrate close to the polymerase active site. The widened minor groove of the primer–template is contacted by active-site residues that assist in positioning the 3'-end of the primer for the nucleotidyl transfer reaction. The fingers must then reopen so that the DNA template can translocate to the next position in the sequence, and a new nucleotide substrate can bind. The DNA template strand does not go through the cleft formed by the fingers and thumb subdomains. Instead, there is a sharp kink at the 5'-end of the template that directs the single-stranded DNA template out of the polymerase active site.

There are many factors influencing reverse transcriptase activity. Some factors that might affect the enzymatic activity of the RT are for example processivity, fidelity, thermostability, and solubility.

Processivity is defined as the number of nucleotides a polymerase is able to incorporate into a strand per single binding event [42]. It could also be described as the probability of not terminating the transcription at a given position. Processivity is closely related to a high catalytic rate and fidelity of the synthesis. Thus, a decrease of polymerase activity likely means a loss of processivity as the enzyme slows down its movement on the template. In general, RTs have lower levels of processivity than other polymerases. The processivity of wtMMLV and RNase H truncated MMLV are 20-40 bases under a single binding event in the presence of a heparin trap [42]. However, given longer reaction times and without a trap, the RTs are able to synthesize templates hundreds of bases long and have been engineered to handle over a thousand bases. By comparison, highly processive polymerases with a sliding clamp such as DNA Pol III, can stay bound for thousands of bases without dissociating [51].

Fidelity is a characteristic defining the enzyme's ability to copy a DNA or RNA template without introducing any errors such as mismatches, deletions or insertions. It also describes an RT's ability to incorporate an error into the nascent DNA strand and still continue the elongation. In contrast to replicative DNA polymerases, RTs does not possess proof-reading activity to reduce the error numbers which gives the RTs relatively low fidelity and more prone to mistakes in comparison to DNA polymerases. In general, RTs have a misincorporation frequency between one in ten

thousands, and one in a million [52][42].

The thermostability of a protein is related to its intrinsic stability, thus, stable proteins will have higher resistance against thermal denaturation [53]. In multiple studies, the affinity to a template has been shown to be closely related to the thermal stability of the RT [42]. The correlation seems to lie in the tighter structure allowing the RT to bind its template more closely. Thermostability affects the overall performance of RTs, defining the ability of RT to synthesize cDNA at elevated temperatures effectively bypassing complex RNA templates structures. For that reason, most studies dedicated to RTs improvement are touching the subject of thermal stability. The same reasoning could also be applied to processivity, as this parameter depends on the tightness of RT-template interactions.

Protein solubility is an important factor for enzyme function. As insoluble proteins tend to misfold or have exposed hydrophobic regions, they often form aggregates leading to a loss of enzymatic activity [54]. In overexpression systems, especially of recombinant proteins, proteins with solubility issues tend to form dense insoluble aggregates called inclusion bodies, due to aggregation [55][54]. RTs in general seem to struggle with solubility and are prone to form inclusion bodies during purification [42].

Aggregation can also be an indication of low intrinsic stability as aggregation tends to occur when hydrophobic regions are exposed during denaturation or folding. Studies have reported a positive correlation between the stability and the rate of folding of a protein [55]. These studies have shown that stable proteins tend to have a higher rate of folding and a lower percentage of exposed hydrophobic residues, making them more resistant to aggregation.

### 2.3.3 Reverse Transcriptase Panel

This section will cover information about each of the RTs that were used in this project. These include the MMLV, Tf1 and Ec48 RT families, along with AstraZeneca proprietary RTs whose sequences and engineering will not be disclosed in this report. The selection of RTs used as references for the study are the following: wild type MMLV (wtMMLV), MMLV PE2, MMLVΔRNaseH, MMLV_MK, MMLV PE6d and bear_MMLV, wtTf1, Tf1 PE6b, Tf1 PE6c, wtEc48, Ec48 PE6a and pol11ERV (sequences in *appendix A.1*). The relation between these RTs are illustrated in *figure 2.6*. The names that will be used for the proprietary RTs in this report are AZRT_1 and its engineered variant eAZRT_1, AZRT_2, AZRT_5, AZRT_11, and AZRT_15.

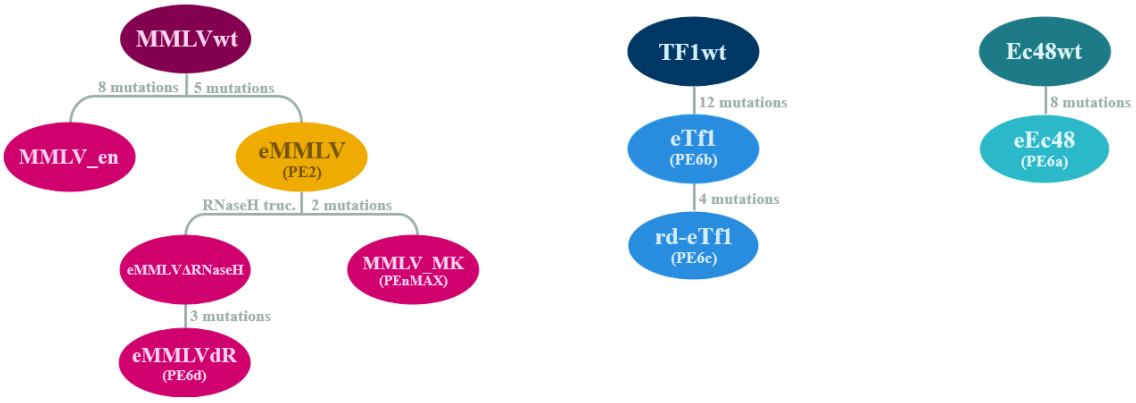
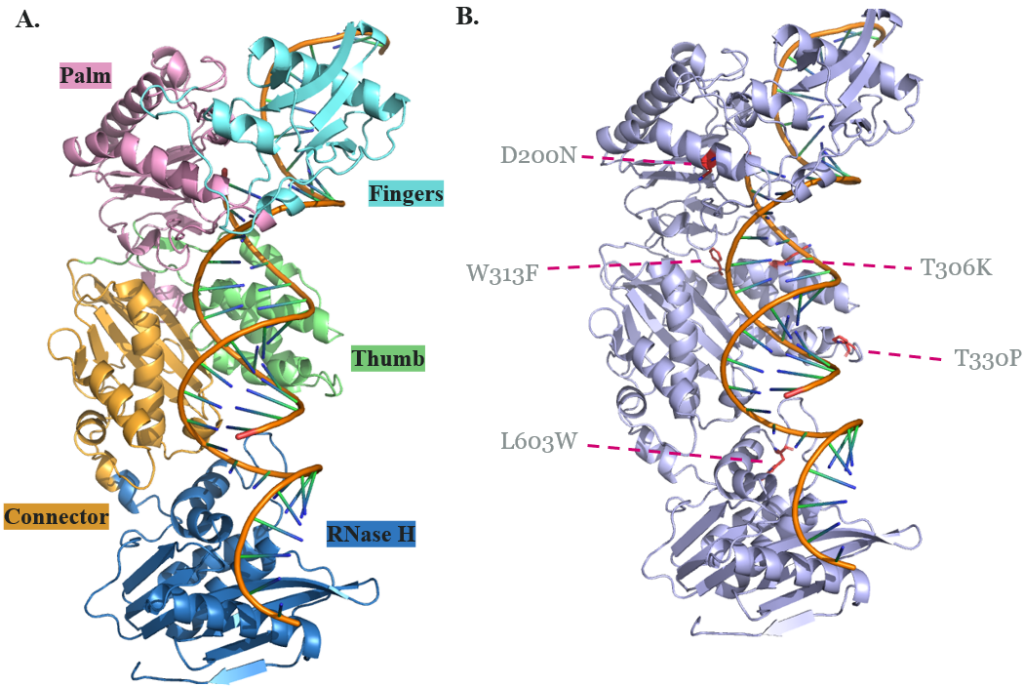



**Figure 2.6:** Family tree of the three RT families MMLV, Tf1 and Ec48.

#### 2.3.3.1 The MMLV RT Family

*Moloney Murine Leukemia Virus* (MMLV) RT is the most widely used RT for TE purposes, and the engineered variant, PE2, is commonly considered the golden standard due to its potency in TE applications. MMLV RT is also quite commonly used for qPCR and other purposes [56][42][57]. MMLV RT has been extensively studied, and a lot of data is available about its properties and performance in different contexts. The structure of MMLV PE2 and its domains and motifs are illustrated in *figure 2.7A and C*.



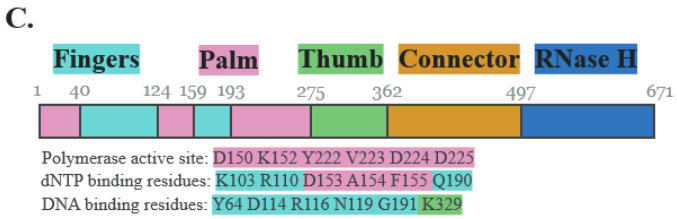



**Figure 2.7:** **A.** Structure and domains of MMLV PE2 (structure predicted with Boltz-2). **B.** The positions of the mutations introduced to MMLV PE2 (structure predicted with Boltz-2). **C.** The domains and important residues of MMLV PE2 [42].

MMLV PE2 is a pentamutant that was created with TE application in mind [9]. A.V. Anzalone et al. that worked on its development, hypothesized that improving thermostability, processivity and DNA/RNA substrate affinity would enhance TE efficiency, which led to the implementation of the five mutations D200N, L603W, T330P, T306K and W313F (*figure 2.7B*). D200N, L603W and T330P was speculated to increase the thermostability of the RT as reported from previous studies, while T306K and W313F supposedly enhanced binding and processivity, as well as thermostability. Compared to wtMMLV, MMLV PE2 installs TEs with significantly higher efficiency and is able to handle a wider range of substrates. The result of the engineering were believed to be consistent with the expected outcomes from improved thermostability and processivity [9].

It has been shown that almost all substitutions of D200 increase thermostability in MMLV RT [42]. However, the residue is close to the polymerase active site, only about 7-9 Å away from the YXDD motif (Y222, V223, D224, D225) (*figure 2.7C*), therefore, substitutions of D200 may cause structural rearrangements of the catalytic and neighboring residues that may alter the substrate-binding affinity and the rate of catalysis.

Residue L603 is positioned at the helix-unstructured loop boundary at the end of the C-helix in the RNase H domain (*figure 2.7B*). All sizeable hydrophobic or aromatic substitutions of position 603 improve thermostability, possibly by stabilizing a hydrophobic Y598-L603-I617 cluster [42]. T330 is located at the terminus of a small helix in the thumb domain. T330P possibly stabilizes the DNA-contacting helical motif. W313F has also been shown to increase the thermal stability of MMLV RT, possibly through improved van der Waals interaction with the nucleic acids. The positive side chain substitution of T306K might enhance nucleic acid binding and thermostability through an increased number of hydrogen bonds with the minor groove of the substrate [42].

The MMLVΔRNaseH has the same set of mutations as the MMLV PE2, though was truncated to minimize the RT by removing the RNase H domain. It has been shown that RNase H deficient MMLV variants are more thermostable than wtMMLV in the presence of substrate due to tighter binding of the template [42]. This means

that both MMLVΔRNaseH and MMLV PE6d, which also has the truncated RNase H domain, should possess higher thermostability in the presence of template than the other variants just based on the domains. However, due to the missing RNase H domain, MMLVΔRNaseH also has lower processivity, speculated to be a result of the loss of the DNA-binding sites of the RNAse H domain [42].

In the case of MMLV PE6d, in contrast to MMLVΔRNaseH, the RNase H domain was truncated as a product of evolution to improve TE efficiency [11]. MMLV PE6d is a further engineered PE2 RT where evolution and rational engineering were combined to to enhance its activity in TE. Three new mutations were introduced in MMLV PE6d, of which one substituted one of the mutations introduced in the MMLV PE2. These three mutations were T128N, V223Y and D200C. The V223Y mutation affects the residue in position X in the YXDD polymerase active site motif and the T128 residue lies close to the polymerase active site and might affect substrate affinity due to structural rearrangements. It is hypothesized that the mutations introduced rescues the lost processivity due to truncation of the RNase H domain. In contrast to MMLVΔRNaseH, that also have a truncated RNase H domain, MMLV PE6d performs well on longer edits and highly structured gRNA that causes the MMLVΔRNaseH to terminate the polymerization prematurely. However, MMLV PE6d seems to mediate an increased number of scaffold insertions compared to, for example, MMLVΔRNaseH. PE6d was determined to be the most efficient and accurate RT for longer, structured RTTs due to its presumed high level of processivity.

MMLV_MK is an MMLV PE2 variant with two additional mutations: V223M and L435K. The RT is designed with the PEmax architecture for increased efficiency. L435K is proposed to improve solubility and thermostability of the RT and might also positively affect affinity to the template [42].

The bear_MMLV is an MMLV variant optimized for qPCR. It was developed in an effort to create an efficient and cost effective RNA extraction- and RT-qPCR kit for SARS-CoV-2 detection during the pandemic [56].

| **RT** | **Mutations** |
|---|---|
| MMLV PE2 | *D200N, L603W, T330P, T306K, W313F* |
| MMLVΔRNaseH | *D200N, L603W, T330P, T306K, W313F,* -RNaseH domain |
| MMLV_MK | *D200N, L603W, T330P, T306K, W313F,* V223M, L435K |
| MMLV PE6d | T128N, V223Y, D200C, *T330P, T306K, W313F,* -RNaseH domain |
| bear_MMLV | Y8H, H204R. M289L, *T306K,* F309N, D524G, E562Q, D583N |

**Table 2.1:** Mutations of MMLV RT variants. Italicized mutations are shared with MMLV PE2.

#### 2.3.3.2 The Tf1 RT Family

Tf1 PE6b (also called evo-Tf1) and Tf1 PE6c (also called rd-evo-Tf1) were engineered from the LTR retrotransposon of fission yeast, *Schizosaccharomyces pombe* (Tf1). The Tf1 is similar to the MMLVΔRNaseH in the sense that it does not possess an RNase H domain (*figure 2.8*).

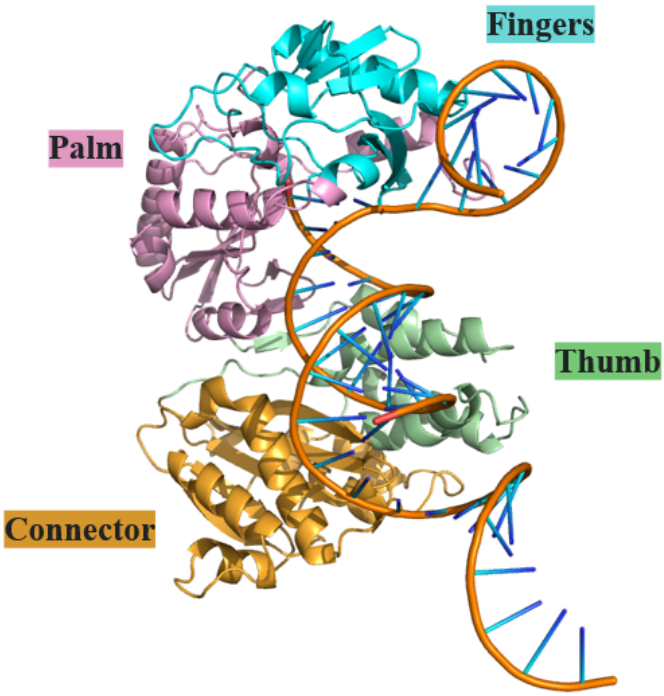



**Figure 2.8:** Structure and approximate domains of Tf1 RT (structure predicted with Boltz-2).

Tf1 PE6b was evolved through phage-assisted evolution by linking the phage propagation to protein activity [11]. The evolution strategy improved RT activity and resulted in improved editing on average 2.7-fold compared to the wild type, and supports TE on equal levels as PE2max in HEK293T cells. However, PE6b shows less efficiency at longer and more complex edits. Several of the resulting mutations (K118R, I128V, K413E, and S492N) are proximal to the DNA/RNA substrate, while others (P70T, G72V, M102I, and K106R) may interact with the RTT of the gRNA. Compared to MMLV PE2, Tf1 PE6b lacks in processivity and is estimated based on performance to have a level of processivity similar to the truncated PE2 [11], however, no evaluation of processivity alone was made.

The final version, Tf1 PE6c, harbors in total 16 mutations. Tf1 PE6c was rationally engineered from the Tf1 PE6b to improve DNA/RNA interaction by introducing five mutations guided by a Tf1 homolog. The PE6c variant possessed improved TE efficiency for longer edits, matching or even exceeding PE2max, and had improved processivity relative to its predecessor. The authors concluded that PE6b is well suited for shorter edits while the high level of processivity that PE6c is speculated

to possess, makes it well adapted for longer, structured RTTs around 60bps where other RTs tend to struggle [11].

| RT | Mutations |
|---|---|
| Tf1 PE6b | *P70T, G72V, S87G, M102I, K106R, K118R, I128V, L158Q, F269L, A363V, K413E, S492N* |
| Tf1 PE6c | *P70T, G72V, S87G, M102I, K106R, K118R, I128V, L158Q, F269L, A363V, K413E, S492N*, S188K, I260L, S297Q, R288Q |

**Table 2.2:** Mutations of Tf1 RT variants. Italicized mutations are possessed by both variants.

#### 2.3.3.3 The Ec48 RT Family

The intention behind the development of Ec48 PE6a RT (also called evo-Ec48) was to develop an editor of significantly smaller size than current existing TEs to facilitate *in vivo* delivery. Ec48 PE6a was therefore engineered from a wild type retron found in *Escherichia coli*, Ec48 RT, approximately 800 bp smaller than the conventional MMLV RT, and with low baseline editing efficiency. Through phage-assisted evolution, 8 mutations emerged that improved TE activity on average 22-fold compared to its wild type counterpart in HEK293T cells [11]. Compared to MMLV PE2 in the PEmax architecture however, PE6a on average only reaches about 80% of the editing efficiency and struggles more with longer and complex edits.

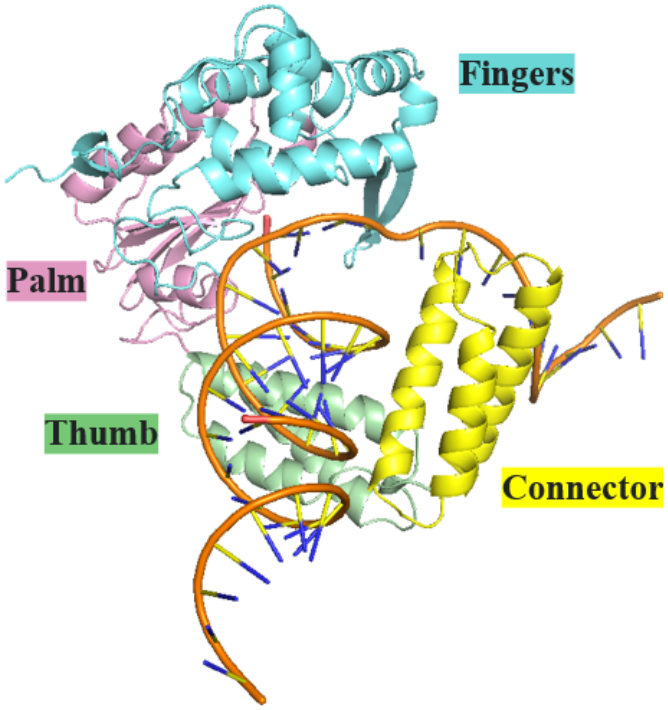



**Figure 2.9:** Structure and approximate domains of Ec48 RT (structure predicted with Boltz-2).

Ec48 RT possess a unique structure in comparison to other established RTs. It has palm, fingers and thumb domain, along with a fourth noncanonical connector domain attached to the thumb domain (*figure 2.9*). The fourth domain has some sequence similarity to the usual RNase H domain in RTs, suggesting an evolutionary rearrangement at one point, however, this is not confirmed. Due to its different structure, it is also possible that its polymerization mechanism is slightly different compared to that of common RTs.

| RT | Mutations |
|---|---|
| Ec48 PE6a | E60K, K87E, E165D, D243N, R267I, E279K, K318E, K343N |

**Table 2.3:** Mutations of Ec48 PE6a RT.

#### 2.3.3.4 pol11ERV RT

The pol11ERV RT has very low intrinsic activity when used in the context of TE [11]. Its structure is similar to that of MMLV RT as it possesses an RNase H domain (*figure 2.10*). It has the size 67 kDa.

The pol11ERV's role in the panel of established RTs is to provide a reference point of bad characteristics of an RT and for possible engineering opportunities.

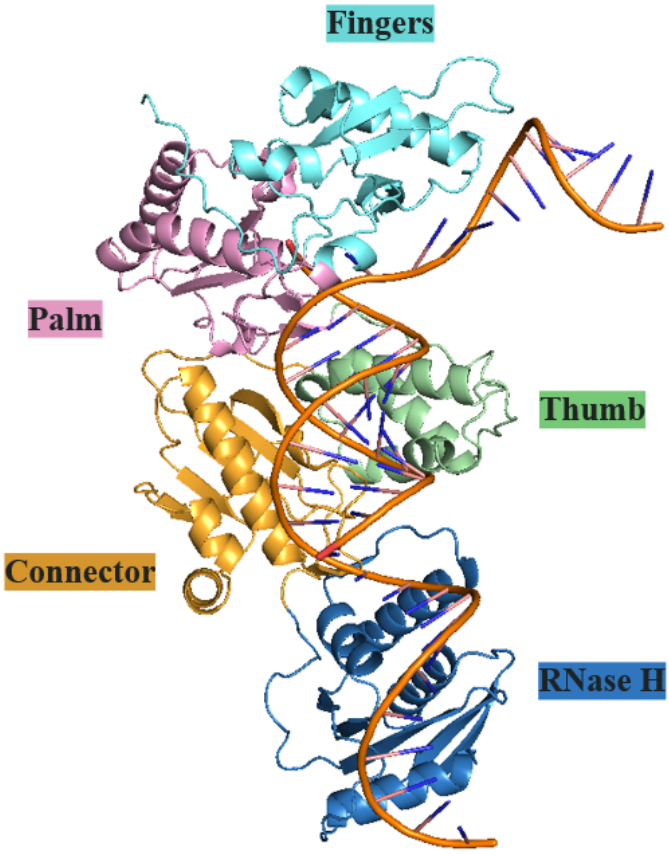



**Figure 2.10:** Structure and approximate domains of pol11ERV RT (structure predicted with Boltz-2).

## 2.4 HEK293T cell line

For this project the immortalized cell line HEK293T was used. The line originates from HEK293 cell line developed from embryonic kidney cells in the 1970s [58]. Since their creation they have been used by the biotechnology industry to produce therapeutic proteins and viruses for gene therapy as well as safety testing of chemicals. HEK293T cells are partially MMR deficient due to hypermethylation of the MLH1 promoter [59]. This explains why the TE efficiency in HEK293T cells are often higher than in other cell lines [9].

## 2.5 Experimental Theory

This section will cover some theory behind the methods used in this project, including system design choices, biochemical and biophysical assays and machine learning tools used for engineering.

### 2.5.1 The Conventional Genome Editing Workflow

Plasmid DNA is a common vector for loading the gene editing system *in vitro*, and has advantages such as stability, modularity and low production costs [60]. The system is usually transfected into the cells using a chemical transfection reagent which overcomes the inherent challenge of introducing negatively charged molecules such as DNA and RNA into cells with a negatively charged membrane, by coating the nucleic acids in a chemical that neutralizes or imparts an overall positive charge to the molecule [61]. For editing of cells, it is recommended that the cells are incubated at 37°C for 48-72 hours before analysis of the editing efficiency [62]. Next generation sequencing (NGS) is the current gold standard for analyzing genome editing efficiency and may be used to analyze off-target effects and confirm gene edits with targeted sequencing [63]. Genome editing experiments result in mixed cell populations due to low efficiency, with likely only a subset carrying the desired edit. Using NGS, both qualitative and quantitative information about the editing outcome can be obtained. Usually, the turnaround time for an experiment from transfection to receiving the results of the NGS analysis is about 2 weeks [64][65], after which significant data handling is needed to process the reads.

### 2.5.2 HiBiT

HiBiT is a small 11 amino acids tag that binds with high affinity to its larger subunit LgBiT [66]. When bound, the complex is able to catalyze a light-emitting reaction upon exposure to the substrate furimazine. By attaching the HiBiT epitope tag to a protein and supplying the cells with LgBiT, the system can be used as a bioluminescent detection method. However, due to LgBiT's large size, it cannot cross the intact cell membrane and requires cell lysis or alternatively, endogenous expression. Luminescence intensity is directly proportional to protein expression in the cells. The luminescence readout from a microplate reader is given in the arbitrary unit "Relative Light Units" (RLU). For comparisons within the same experiment the system is considered quantitative, however, it is not reliable to compare RLU across experiments as the readout is not standardized and relies on photon emission, amplification, and conversion that might vary between every run depending on random factors [67].

### 2.5.3 Reporter System for TE Activity

The reporter system of choice for assessing TE activity was a HiBiT assay previously developed in-house at AstraZeneca [68]. The idea is to tag the protein of interest with HiBiT through TE and measure the luminescence activity as a proxy for editing efficiency (*figure 2.11A*). The relative luminescence activity levels have been shown to correlate well with NGS data. The advantage of this reporter assay is the easy workflow and fast turn-around time of about a week in comparison to NGS.

The system utilizes a split PEn editor and springRNA guide designed to tag the C-terminal end of a highly expressed gene of interest with HiBiT and a short linker, followed by a stop codon to terminate translation (*figure 2.11*B). The insertion is

approximately 50 bp in total. The TE consists of two plasmids, one containing an RT and one containing wild type nuclease SpCas9 and GFP. The advantage of the split editor system is the modularity. The RT component can be switched out without changing the Cas9 plasmid which also increases the cost-effectiveness. The GFP on the Cas9 plasmid is expressed at equimolar ratio to the SpCas9, through ribosomal skipping encoded by the T2A sequence. The GFP expression provides a reliable proxy for the transfection efficiency of the Cas9 plasmid which is used for normalization of the readout together with cell confluence. The two genes of interest (GOI) for tagging, were the highly expressed GAPDH and MIF1 genes. PAM sites for editing were chosen near the terminal end of these genes to insert the protein tag and associated linker in-frame (*figure 2.11*C). For the MIF1 site, orientation of the optimal spacer was to target the antisense strand, meaning that longer-than-intended insertions could potentially create out-of-frame insertions of the HiBiT tag, which would decrease luminescence output [68].

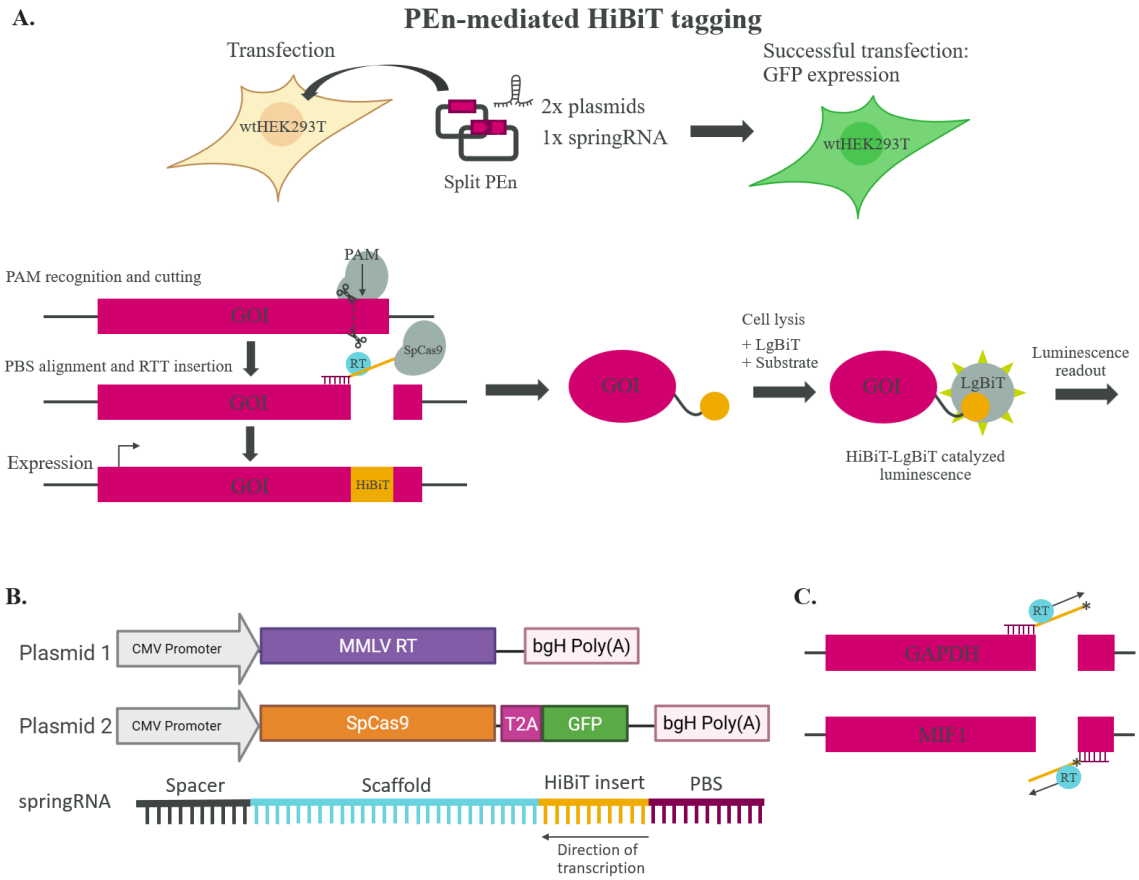



**Figure 2.11:** **A.** PEn-mediated HiBiT tagging reporter system.**B.** Components of the split PEn editor. **C.** Direction of insertions at GAPDH and MIF1 genes respectively.

To be able to screen RTs reliably while minimizing the influence of other unknown factors, all components of the editor are overexpressed. This method aims to saturate the system and make the RT component the determining factor of editing

efficiency. This was achieved by putting both the RT and the Cas9 under the regulation of the CMV promoter which induces a very strong expression of the proteins in HEK293T cells. The cells were also supplied with an excess of gRNA.

#### 2.5.3.1 Target Sites

Two loci of interest in human cells are GAPDH and MIF1. The GAPDH gene encodes a member of the glyceraldehyde-3-phosphate dehydrogenase protein family [69]. The protein is highly expressed and catalyzes an important energy-yielding step in carbohydrate metabolism and is essential for cell survival. According to the Human Protein Atlas, the RNA expression of GAPDH in HEK293 cells is approximately 3000 nTPM (normalized transcripts per million) [69].

Macrophage Migration Inhibitory Factor 1 (MIF1), encodes a lymphokine involved in the immune response. It plays a role in regulating macrophage function, however, is not essential for cell survival. According to the Human Protein Atlas, the RNA expression of MIF1 in HEK293 cells is approximately 4600 nTPM [69].

For reference, other housekeeping genes such as $\alpha$-tubulin and $\beta$-actin (ACTB) are expressed at levels of about 130[70] and 1570[71] nTPM respectively in HEK293 cells, according to the Human Protein Atlas. The extremely highly expressed gene Eukaryotic translation elongation factor 1 alpha 1 (EEF1A1) is expressed at levels of around 6700 nTPM in HEK293 cells [72].

### 2.5.4 Promoters

Ten previously studied promoters were selected based on their reported strength, to obtain a panel of promoters from weak to strong with multiple intermediates (sequences in *appendix A.2*). There is no study that compares all promoters simultaneously, but one study covers 7 out of the 10 promoters in HEK293T cells [73]. The estimated order of strength of these 7 promoters is as follows (from strongest to weakest): CMV, EF-1$\alpha$, PGK, EFS, UbC, SV40, HSV-TK [73][74]. The three remaining promoters are JeT, GAPDH and EIF4A1. The JeT promoter is predicted to have low to medium expression [75]. GAPDH is predicted to have strong expression [76]. EIF4A1 is predicted to have medium to strong expression [77].

It has been shown that the kinetics of the CMV promoter is very fast, it gives strong expression of the transgene immediately, and reaches a peak in expression within 24 hours, then the expression level slowly decreases with time [78][79]. In contrast, the EF-1$\alpha$ promoter is stable for long-term expression and takes about 72 hours to reach peak, maximum expression [80][81]. Within 24 hours the expression level is about 90% of maximum expression. The protein levels are high from this promoter. The PGK promoter is also relatively strong in mammalian cells, though slow to reach maximum expression. After about 72 hours, PGK have reached peak expression and keeps a stable expression over time [82][81]. The UbC promoter takes about five days to reach maximum expression post-transfection, but keeps a relatively stable expression over time after that [79]. EFS, which is a shorter version

of EF-1$\alpha$, also exhibits long-term stable expression [83]. In contrast to its longer version, has very high initial expression, in some cell lines even exceeding CMV expression at 24h.

### 2.5.5 Fluorescence Polarization

Fluorescence polarization (FP) is a method that uses polarized light to measure the rotation of a molecule between the time of absorption and emission of fluorescence [84]. The principle behind the technique is that the light emission of a fluorescent probe attached to a small molecule, will change if the small molecule interacts with something larger. Rapid tumbling of the molecule will cause the emitted light to become depolarized when excited by a laser, but a bound ligand will cause the emitted light to stay polarized due to the slowed tumbling. The fluorescence polarization of the light can be quantified to determine the proportion of bound proteins. FP is given in the standard unit millipolarization (mP). Higher mP values mean slower movement, which is an indication of binding. FP data can be used to obtain the equilibrium dissociation constant, Kd, which is the protein concentration necessary to bind half of the available substrates at equilibrium. A smaller Kd indicate a strong binding interaction between enzyme and ligand. To obtain Kd, the concentration of RT is titrated against a constant concentration of labeled substrate.

The Hill model is used to describe single binding events and fits a sigmoidal curve [85]. The Hill equation is described by the following formula: $Y = \frac{Bmax \times X^h}{(Kd^h + X^h)}$ where Y is binding in mP, X is the protein concentration, Kd is the protein concentration necessary to bind half of the available substrate at equilibrium, Bmax describes the maximum specific binding in the unit mP, and h is the Hill constant used to describe the level of cooperativity in the binding. Hill coefficient equals 1.0 when a monomer binds with no cooperativity to one site.

### 2.5.6 Thermal Shift Assay

A thermal shift assay can be used to determine a protein's melting point (Tm) by tracking how its physical properties changes during heating [53]. In thermal shift assay a fluorescent dye is added to the protein mixture, and as the protein denatures with the increasing temperature, the dye will interact with the hydrophobic core of the proteins that are being exposed. The dye will release a fluorescent signal when it is bound to the protein, resulting in a melting curve. The protein's melting point is the temperature at which the melting curve is the steepest.

### 2.5.7 AI Tools for Protein Prediction and Engineering

#### 2.5.7.1 AlphaFold2

AlphaFold2 is a computational tool that utilize machine learning to predict 3D protein structures from an amino acid sequences [86]. AlphaFold was developed by DeepMind and is capable of predicting protein structures to near experimental accuracy in a majority of cases with a backbone accuracy of 0.96 Å. AlphaFold

greatly improves the accuracy of structure prediction compared to competitors by incorporating novel neural network architectures and training procedures based on the evolutionary, physical and geometric constraints of protein structures [86]. It outputs per-residue coordinates and confidence metrics pLDDT (per-residue local confidence) used to determine the reliability of the prediction. AlphaFold2 embed multiple sequence alignments and pairwise features to enable structure prediction, and the predictions are refined through iteration. The network directly predicts 3D atomic coordinates from a protein's primary amino acid sequence and aligned homologous sequences.

#### 2.5.7.2 Boltz-2

Boltz-2 also predicts protein structures through multiple sequence alignments [87]. It extracts evolutionary couplings that inform residue contacts, and employs iterative refinement to improve predictions. Its outputs include backbone atom coordinates along with confidence estimates of prediction reliability. Boltz-2 can also predict biomolecular complex structures. The accuracy of the predicted interface is assessed using ipTM (interfacial predicted TM-score), a confidence metric that estimates the reliability of between-chain contacts. Similar to AlphaFold, Boltz-2 has been pre-trained on experimentally solved structures from the PDB and large sequence resources.

#### 2.5.7.3 PyMOL

PyMOL is a molecular visualization and analysis tool for viewing, editing, and rendering 3D structures of proteins, nucleic acids, and small molecules [88]. Users can interactively explore molecular geometry, measure distances and angles, color by properties, and create cartoons, surfaces, and electrostatic maps. PyMOL also includes a Python-based command line and is capable of structure alignments, superposing of structures, and mutagenesis [89].

#### 2.5.7.4 ProteinMPNN

ProteinMPNN is a graph neural network for protein sequence design that, given a fixed protein backbone structure, predicts amino acid sequences that are compatible with that structure [90]. ProteinMPNN will represent a protein as a graph of interacting residues and uses message-passing neural networks (MPNN) to recover an amino acid sequence from information about backbone coordinates. Graph neural networks will compute independent, concurrent tasks (nodes) simultaneously [91]. To synchronize data across multiple processes they utilize message-passing to update information from parallel processes. Message-passing involves exchange and aggregation of information from neighboring nodes.

In ProteinMPNN each coordinate represents a node, which has information about itself such as its coordinate, orientation and distance to neighbors [90]. In each step, each node reads notes from its neighboring nodes, combines them, and then updates its own state based on what it learned. Repeating this for several layers lets

information propagate through the network so even far-apart points indirectly share context. Through successive message-passing layers, each residue's representation is updated by aggregating information from its neighbors, so the node's state comes to reflect constraints like packing, hydrogen-bonding geometry, and steric compatibility. After several rounds, the model will process the information, and for each coordinate select an amino acid that best satisfy the learned structural constraints at that site through a scoring distribution reflecting each amino acid's fit.

Pre-training of ProteinMPNN included supervised learning on structures from the PDB [90]. The network was optimized to emphasize patterns that correlate with successful side-chain packing and fold stabilization by encouraging it to assign high probability to the true residue at each position. The performance was further improved by training with backbone noise and interference that contain less details of the backbone geometry. The overall sequences recoveries of native proteins from the test set were about 52%.

A version of ProteinMPNN called soluble ProteinMPNN was retrained on soluble proteins to provide sequence prediction that will yield soluble structures [92]. Soluble ProteinMPNN will reduce hydrophobic residues on the surface while preserving the target topology.

#### 2.5.7.5 ThermoMPNN

ThermoMPNN is a graph neural network model for protein thermostability prediction using the MPNN communication model to inform its predictions (thermo for thermostability) [93]. ThermoMPNN takes an amino acid sequence and protein structure as input and in the network, each residue will represents a node with information such as amino acid type, secondary structure, contacts and orientation. Message-passing allows ThermoMPNN to predict what amino acids are likely to form favorable interactions if placed at a given position through a combination of structure and sequence information [94]. ThermoMPNN uses sequence and node embeddings along with amino acid interactions from ProteinMPNN to inform its predictions.

Using the message-passing communication model, ThermoMPNN will estimate how mutations to protein sequences will affect stability proxies such as Gibbs free energy ($\Delta$G). An extension of ThermoMPNN called ThermoMPNN-D will also predict the stability changes due to double point mutations by evaluating a pair of double mutations jointly and separately [94]. ThermoMPNN was trained by the developers on datasets with experimentally measured stability readouts and double mutants from available datasets.

# 3

# Methods

This chapter will cover the protocols for the methods in this project and explain protein engineering approaches.

## 3.1 Transformation and Plasmid Maxiprep

0.5 µL (50 ng) plasmid was mixed with 5 µL of competent Dh5$\alpha$ E.coli cells (Thermo Scientific, Cat.#EC0112) on ice. Cells were heat shocked at 42° in a heat block for 45-60 seconds then put back on ice. Cells were mixed with 500 µL LB media and allowed to recover for 30 minutes at 37°. Carbenicillin was mixed with terrific broth (TB) with 0.4% v/v glycerol, to the ratio 1:1000 (antibiotic:media). TB Media was heated to 37° while cells were recovering. 30 mL of TB+carbenicillin was added to flasks and all the transformation mixture was inoculated into the fresh media. The cultures were incubated at 37° at 222 rpm for approximately 16h before harvest. Cultures were transferred to 50 mL falcon tubes before harvest. The cells were harvested and the plasmids extracted following the protocol included with the QIAGEN Plasmid Plus Maxi Kit (Qiagen, Cat.#12963) [95]. Deviations from the protocol included double washing of the DNA with ETR and PE buffer. Plasmids were eluted in 200 µL elution buffer included in the kit, and DNA concentration measured using nanodrop. All plasmids were standardized with elution buffer to a working concentration of 100 ng/µL before use.

### 3.1.1 Mammalian Cell Culture Maintenance

HEK293T cell cultures were split two times a week. Cells were cultured in DMEM (1X) + GlutaMAX (gibco, Cat.#31966-021) and 10% FBS (Thermo Scientific, Cat.#A5256701). Plates with between 80-95% confluence were used for passing. Whole procedure was performed in a laminar air flow (laf) bench.

When splitting cells, the media was equilibrated to 37°C in a bead bath prior to usage. Media was aspirated using a vacuum pump connected to a collection bottle with a narrow clean stripette tip. 7 mL PBS (gibco, Cat.#10010-015) was carefully added to the plate without disturbing the cell layer to wash the remaining media from the cells. The PBS was aspirated using the vacuum pump. Then 3 mL of

TrypLE Express without pH indicator (gibco, Cat.#12604-013) was added to the cells. The cells were incubated with the TrypLE Express at room temperature for approximately 5 minutes until the cells had detached. The TrypLE Express was quenched by addition of 7 mL or the DMEM + FBS media. The media was mixed with the cells to a homogeneous mixture and added to a 50 mL falcon tube. The cells were pelleted through centrifugation at 15.9 x g for 3 minutes. Then the media was aspirated and 1 mL of fresh media was added to resuspend the cells using a 1000 µL pipette. The cells were mixed well with the media by pipetting to separate clumps of cells and then diluted with DMEM + FBS media to a final volume of 10 mL. 12 mL DMEM + FBS media were added to 10 cm plastic mammalian culture dishes, whereafter 0.25-0.5 µL cell suspension was added to each dish. The cell concentration depended on when the culture was needed for experiment. The culture dishes were gently shaken to spread out the cells evenly in the dish and then incubated at 37°C with 5 %$CO_2$, 20% $O_2$ and until further use. Plates were thrown away after a week without use.

### 3.1.2 HiBiT Lysis Buffer

Mammalian lysis buffer was prepared in 1x aliquots and consisted of 20 mM Hepes pH 7.5 (Thermo Scientific, Cat#J60712), 100 mM KCl (Invitrogen, Cat.#AM9640G), 5% glycerol (Thermo Scientific, Cat.#17904), 5 mM $MgCl_2$ (Invitrogen, Cat.#AM9530G), 0.1% Triton-X100 (Sigma-Aldrich, Cat.#T8787-50ML), SIGMAFAST protease inhibitor cocktail tablet (Sigma-Aldrich, Cat.#S8830-20TAB), 1 mM DTT.

### 3.1.3 Transfection and Readout Manual Workflow

Cells were split at approximately 90% confluence and counted using a cell counter. 12,500 cells were seeded per well in a 96-well, black, flat bottomed plate with collagen type I coating (Greiner, Cat.#655956). The plate was incubated at 37°C with 5% $CO_2$ and 20% $O_2$ for 24 hours before transfection.

For transfection of one well with 12,500 cells: Plasmids were added to a PCR strip in a ratio of 1:1 with 50 ng (0.5 µL) of each plasmid. The PCR strip was closed, and spun down quickly using a table top centrifuge. For all experiments, the empty plasmid pUC19 was used instead of the RT component as a negative control. It was co-transfected with the Cas9 plasmid to make the transfection conditions as similar as possible to the usual transfections.

**The 3-component system:** A master mix solution of gRNA and Opti-MEM reduced serum (Gibco, Cat.#31985070) with final concentration of 1 µM (1 pmol/µL) was prepared. 2 µL of the gRNA master mix was added to each PCR tube. 16.6 µL Opti-MEM was added to each tube, then 0.4 µL (1:0.4 DNA:X-tremeGENE) X-tremeGENE DNA transfection reagent (Roche, Cat.#6365779001) that was briefly vortexed to a final volume of 20 µL. The transfection reagent was incubated with the DNA for 10-15 minutes to form complexes before the total volume of 20 µL was gently added to the well by tilting the plate and adding the transfection reagent on

the side of the well. The plate was incubated at 37°C with 5 %$CO_2$, 20% $O_2$, 95% rH for 72 hours before readout.

The plate was photographed using Incucyte (Sartorius) with phase and green filter, and the phase area along with GFP area was calculated using the software. Then the media in the wells were aspirated using a vacuum pump connected to a collection bottle with a narrow clean stripette tip. Then 50 µL of lysis buffer was added to each well. The plate was incubated at room temperature for approximately 3 minutes, shaking the plate for about a minute to facilitate the cell detachment and lysis. The detection reagent was prepared by mixing LgBiT protein from the Nano-Glo Hi-BiT Lytic Detection System (Promega, Cat#N3030) with HiBiT substrate from the same kit (1ml:10µL and 1ml:40µL respectively) in PBS (Gibco, Cat.#10010015). 50 µL of the detection reagent was added to each well. The plate was then incubated at room temperature for approximately 10 minutes before readout in a BMG LABTECH PHERAstar FSX with integrated luminescence detection module. Gain was adjusted to the well with highest level of luminescence, usually resulted in gain 3600. The data was analyzed in Graphpad Prism.

**The 2-component system:** Opti-MEM reduced serum (Gibco, Cat.#31985070) was added to each tube, then 0.6 µL FuGENE HD Transfection Reagent (Promega, Cat.#E2311) was added to each tube, final volume in the tube 20 µL. The transfection reagent was incubated with the DNA for 10-15 minutes to form complexes before the total volume of 20 µL was gently added to the well by tilting the plate and adding the transfection reagent on the side of the well. The plate was incubated at 37°C with 5% $CO_2$, 20% $O_2$, 95% rH for 72 hours before readout.

The plate was photographed using Incucyte (Sartorius) with phase and green filter, and the phase area along with GFP area was calculated using the software. Then the media in the wells were aspirated using a vacuum pump connected to a collection bottle with a narrow clean stripette tip. Then 100 µL of Opti-MEM was added to each well. The live detection reagent was prepared by mixing HiBiT substrate from the Nano-Glo Live Cell Assay System (Promega, Cat.#N2012) with the live detection buffer from the same kit in the ratio 20µL:1ml. 25 µL of the live detection reagent was added to each well. The plate was then incubated at room temperature for approximately 10 minutes before readout in a BMG LABTECH PHERAstar FSX with integrated luminescence detection module. Gain was adjusted to the well with highest level of luminescence, usually resulted in gain 3600. The data was analyzed in Graphpad Prism.

The RLU is normalized to the transfection efficiency using the formula below, unless otherwise stated.

$$\frac{RLU}{GFP/Phase\,area\,(\%)/100}$$

### 3.1.4 Transfection and Readout Automated Workflow

This workflow was only applied for the 2-component system. Plasmids were added to Echo Mass Spectrometry (MS) Qualified 384-well Polypropylene (PP) Microplate (Beckman Coulter, Cat.#C74290) defined as the source plate. Transfer files were prepared in excel and saved as .csv files. Echo655 Liquid Handler were used to transfer a total of 100 nL of each plasmid to the correct wells on a black, 384-well, PS, μclear, sterile, destination plate (greiner, Cat.#781091) according to the transfer file.

The destination plate was centrifuged for 1 minute at 18.4xg to collect plasmid in the bottom of the well. Cells were split at 80-95% confluency, counted using a cell counter and diluted in DMEM (1X) + GlutaMAX (gibco, Cat.#31966-021) and 10% FBS (Thermo Scientific, Cat.#A5256701) to final concentration 300,000 cells/mL. Using a Dragonfly Liquid dispenser (sptLabtech) 40 μL PBS (gibco, Cat.#10010-015) was added in each well of the two outermost rows on all four sides of the plate to minimize edge-effect. FuGENE HD Transfection Reagent (Promega, Cat.#E2311) was mixed in Opti-MEM reduced serum (Gibco, Cat.#31985070) to final concentration of 1:6 (μg DNA to μL FuGENE HD). 5 μL of the FuGENE + Opti-MEM transfection mixture was added to each sample well using Dragonfly and incubated at room temperature for 10-15 minutes. Thereafter, the 300,000 cells/mL mixture was mixed and 25 μL was added to each sample well. The plate was spun down for 1 minutes at 18.4xg and then incubated at 37°C with 5%$CO_2$, 20% $O_2$, 95% rH for 72 hours before readout.

Confluence and GFP expression were studied using Incucyte (Sartorius) and Phase area and GFP intensity were calculated for each well. Nano-Glo Live Cell Assay System (Promega, Cat.#N2012) was used for the readout. The NanoGlo substrate from the kit was mixed with NanoGlo Live Assay buffer to the ratio 1:20 (μL substrate to mL buffer). This substrate mix was added to HEPES (gibco, Cat.#15630-056) and PBS to the final concentration of 80% substrate mix, 10% HEPES (25 mM) and 10% PBS. 10 μL of the prepared reagent was added to each sample well using Dragonfly. Then incubated for 10 minutes before readout in a BMG LABTECH PHERAstar FSX with integrated luminescence detection module. Gain was adjusted to the well with highest level of luminescence, usually resulted in gain 3600. The data was analyzed in Graphpad Prism.

## 3.2 Protein Purification

Protein purification was performed with assistance from another department at AstraZeneca. The process included recombinant expression in *E. coli*, extraction, MS verification and size exclusion chromatography. The buffer used to store the proteins contained: 20 mM HEPES, 250 mM NaCl, 1 mM TCEP, 10% glycerol, 0.01% Tx100 and 0.1 mM EDTA (unknown catalog numbers as the buffer was supplied already mixed with the proteins to me).

## 3.3 Biochemical Assays

### 3.3.1 Substrate Annealing

Substrates were annealed prior to inclusion in biochemical assays. The annealing buffer was prepared in 2x aliquots and was composed of 20 mM TRIS-Cl pH 7.5 (Invitrogen, Cat.#15567027), 100 mM KCl (Invitrogen, Cat.#AM9640G), 2 mM EDTA (Invitrogen, Cat.#AM9260G). To prepare a 10 μM stock of annealed substrate, the labeled DNA primer was mixed with the unlabeled RNA template in 1:1.1 ratio with final concentration of primer being 10 μM and template slightly more to make sure all primer would be annealed. The oligos were mixed with 2x annealing buffer and MQ to a final concentration of 1x annealing buffer. The mixtures were incubated at 95°C for 5 minutes, then slowly cooled to room temperature whereafter stored at -20°C.

### 3.3.2 Reaction Buffer

The reaction buffer used in the biochemical assays were prepared in 5x or 2x or 1x aliquots, and composed of (at 1x concentration) 20 mM TRIS-Cl pH 7.5 (Invitrogen, Cat.#15567027), 100 mM KCl (Invitrogen, Cat.#AM9640G), 5% glycerol (Thermo Scientific, Cat.#17904), 1 mM DTT.

### 3.3.3 DNA/RNA Binding

To assess binding interaction between DNA/RNA complexes and RT, a fluorescence polarization (FP) assay was deployed. 6μL 2x reaction buffer was added to each well in a 384 small volume wells, non-binding, PS, black assay plate (Greiner, Cat.#784900), except in row A. 12μL 2x RT solution in 2x reaction buffer was added to each well in row A, see *table 3.1* for the concentrations of each protein mix. With a 16-channel pipette, 6μL of the protein solution was aspirated and added to the second row on the plate. Protein solution was mixed with the fresh buffer 20x times to create a homogeneous 2x dilution of the protein. The procedure was repeated for each row, creating a 2x dilution series, skipping the last row. 3 μL, 4 nM annealed DNA/RNA substrate (DNA primer with 5' FAM fluorophore), either GAPDH_DNA_primer with GAPDH_RNA_template_blunt or MIF1_DNA_primer with MIF1_RNA_template_blunt or any combination of MIF1 RNA or DNA oligos or ssDNA or ssRNA (sequences in *table A.1*), and 3μL 4μM $MgCl_2$ was added to all wells. The plate was shaken briefly to mix the components and then incubated for 30 minutes at room temperature. Fluorescence was read in BMG LABTECH PHERAstar FSX with Optic Module FP 485 520 520. Gain and focus were adjusted to a well without protein to create a target baseline of 35mP. The data were analyzed in Graphpad Prism and fitted to the Hill model to obtain Kd.

| RT | Final Concentration [µM] |
|---|---|
| wtMMLV | 9 |
| MMLV PE2 | 10 |
| MMLVΔRNaseH | 12 |
| MMLV_MK | 10 |
| MMLV PE6d | 14 |
| bear_MMLV | 15 |
| wtTf1 | 6 |
| Tf1 PE6b | 12 |
| Tf1 PE6c | 8 |
| wtEc48 | 12 |
| Ec48 PE6a | 8 |
| AZRT_1 | 13 |
| eAZRT_1 | 14 |
| AZRT_2 | 15 |
| AZRT_5 | 2.5 |
| AZRT_11 | 10 |

**Table 3.1:** Table of individual concentrations for RTs in FP assay. Protein stock concentration was the limiting factor of the reaction concentrations.

### 3.3.4 Activity Assay

To assess the polymerization activity of the RTs a polymerization reaction was conducted and resolved through denaturing UREA PAGE. RT protein stocks were mixed with 5x reaction buffer to a 4x master mix of 4 µM RT and 1x reaction buffer. Annealed substrate was mixed with 5x reaction buffer to a 4x master mix of 80 nM annealed substrate (either GAPDH_DNA_primer with GAPDH_RNA_template_blunt or MIF1_DNA_primer with MIF1_RNA_template_blunt, sequences in *table A.1*) and 1x reaction buffer. $MgCl_2$ (Invitrogen, Cat.#AM9530G) was mixed with 5x reaction buffer to a 4x master mix of 4 mM $MgCl_2$ and 1x reaction buffer. dNTP mixed nucleotides (Thermo Scientific, Cat.#R0192) was mixed with 5x reaction buffer to a 4x master mix of 400 µM dNTP mix and 1x reaction buffer. A stop solution for quenching of the polymerization reaction was prepared by mixing 0.5M EDTA (Invitrogen, Cat.#10135423) with Proteinase K (Thermo Scientific, Cat.#EO0492) in the ratio 1:1. 3 µL of the 4x RT mix, 4x $MgCl_2$ mix, and the 4x substrate mix was mixed in a PCR strip. 3 µL of the 4x dNTP mix was prepared in another strip, 1.5 µL of the 6x stop solution was prepared in a second separate PCR strip. The strip containing RT+substrate+$MgCl_2$ and the strip containing dNTPs were incubated in 37°C for 5 minutes in a Biorad thermocycler. Using a

multichannel pipette, the RT+substrate+$MgCl_2$ mixture was added to the dNTP mixture, mixed through pipetting and incubated for 20 seconds. At 20 seconds the RT+substrate+$MgCl_2$+dNTP mixture was added to the strip containing stop solution and quickly mixed through pipetting to quench the polymerization. Thereafter, the reaction mixtures were incubated at 54°C for 30 minutes. In fume hood, HI-DI formamide+orange G was added to each PCR tube to a ratio of 1:1. The mixtures were then boiled at 98°C for 10 minutes before loading on a Novex™TBE-Urea Gels, 6% (Invitrogen, Cat.#EC68755BOX) in fume hood. The gel was run at 90V for 15 minutes, then 200V for 25 minutes. The gel was put on a black tray and studied in a Biorad ChemDoc Imaging System with the program ALEXA647.

As negative control, a tube without protein was prepared and handled identically to the samples. Instead of protein, the same volume of 1x reaction buffer was used.

As a positive control, and reference for what a completed reaction was, ThermoScientific's Multiscribe RT from the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Cat#4368814) was used and the reaction prepared according to the user guide [96]. The volumes were scaled to 6 µL total.

To calculate percentage extension, the gel pictures were processed in ImageJ. Each lane was selected and intensity of the bands were plotted as peaks (*figure B.6-B.12*). A line was drawn from the base of the 70 bps extension (the intended product) to the base of the unextended primer. ImageJ was used to calculate the area of these peaks. Any peaks sticking up between the fully extended product and the primer were included in the total amount of DNA, however, longer fragments than the completed product were disregarded as annealed residuals. The area of the peak corresponding to the intended product was divided on the total area of all the aforementioned peaks to obtain the percentage (*figure B.4*). For all positive controls on each gel the same was done. Mean was calculated for each sample and positive controls. The samples were then normalized to the positive control by dividing the mean of the samples with the mean of the positive controls on the same gel.

### 3.3.5 Nuclease Activity

To determine RT nuclease activity, RTs were incubated together with 20nM annealed DNA/RNA substrate (DNA primer labeled with ATTO647 fluorophore) and 1uM $MgCl_2$ (Invitrogen, Cat.#AM9530G) at 37°C for 2h, total volume 6uL. Next, the proteins were degraded by Proteinase K (Thermo Scientific, Cat.#EO0492) at 54°C for 30 minutes. HI-DI Formamide with Orange G was added to a volumetric ratio of 1:1 v/v. Samples were boiled at 98°C for 10 minutes and 3uL was loaded into each well on a Novex™TBE-Urea Gels, 10% (Invitrogen, Cat.#EC68755BOX) and run at 90V for 15 minutes, then 200V for 25 minutes. The gel was put on a black tray and studied in a Biorad ChemDoc Imaging System with the program ALEXA647.

## 3.4 Thermostability

Thermostability was measured through thermal shift assay with the fluorescent dye SyPRO Orange (Invitrogen, Cat.#S6651). The proteins were diluted with buffer containing KCl (Invitrogen, Cat.#AM9640G), $MgCl_2$ (Invitrogen, Cat.#AM9530G), HEPES pH 7.5 (Invitrogen, Cat.#J60712), and glycerol (Thermo Scientific, Cat. #17904) to the final concentration 0.1 mg/mL protein, 150 mM KCl, 1 mM $MgCl_2$, 20mM HEPES, 2% glycerol. 50 mM of the SyPRO Orange dye was added to each well in a 384-well PCR microplate for Roche 480 LightCycler (Corning, Cat. #AXYPCR384LC480WNF). Four replicates, along with the blanks with only buffer. 10 µL 0.1 mg/mL protein was added to each well and mixed 20 times with a pipette. The plate was sealed with Roche LightCycler sealing foil (Roche, Cat.#04729757001). Plate was shaken vigorously for 1 minute, then spun down briefly using a plate centrifuge. The plate was analyzed in Roche LightCycler 480 with a thermocycler protocol heating the plate from 20 to 95°C with 0.02 degrees increase per minute. Heating curves were plotted and Tm calculated by finding the maximum of the derivative curve.

## 3.5 Aggregation Temperature and Tryptophan Fluorescence

The aggregation temperature was measured through turbidity of the protein solutions during heating. Simultaneously tryptophan fluorescence was measured through 350/330 nm ratio emission spectra, resulting in melting curves. Stock solutions of proteins were thawed rapidly and put on ice. Proteins were centrifuged in a table top centrifuge at 15,000xg at 4°C for 10 minutes. Approximately 10 µL protein sample were loaded into glass capillaries and put into a Promethus Panta instrument (Cornell Scientific) for characterization, along with a blank with buffer only. Samples were heated up from 25°C to 95°C and turbidity and 350/330 nm emission were measured by the instrument throughout the heating with increment of 0.1 degree per minute.

## 3.6 Western Blot

Cells were split and counted in cell counter, then 1.2 million cells was seeded per well in 6-well Corning Falcon Cell Culture Plate (Corning, Cat.#CLS353046). The next day, each well was transfected with 100 µL transfection reagent containing 16 µL plasmid encoding MMLV PE2 RT linked to a 3x FLAG tag and negative control without the tag, 9.6 µL Fugene HD Transfection Reagent (Promega, Cat.#E2311) and 74.4 µL Opti-MEM (Gibco, Cat.#31985070). The plate was incubated at 37°C in 5% $CO_2$, 20% $O_2$ for 72 hours. Day before harvest of the cells, PBS (gibco, Cat.#10010-015) was put in the fridge at 4°C. After 72 hours, the cells were washed twice with cold PBS by adding the PBS gently to the cell wall and removing with a 1000 mL pipette. The cells were left bare without media and put on ice. RIPA

buffer (Thermo Scientific, Cat.#89901) was prepared in 1.5 mL tube by mixing 1 mL RIPA buffer with 1 µL Halt protease inhibitor cocktail (Thermo Scientific, Cat.#1862209). 150 µL of the RIPA buffer was added to each well. A scraper was used to scrape cells from the wells. The cells were put into 1.5 mL pre-cooled tubes. The tubes were spun down at 11,000xg for 10 minutes at 4°C in a table top centrifuge. The supernatants were transferred to a fresh 1.5 mL pre-cooled tube.

To standardize the protein concentrations for the western blot, a BCA assay was performed according to ThermoScientific Pierce BCA Protein Assay Kit [97] (Thermo Scientific, Cat.#23235). 5 µL sample and standard were added to a 96 well corning assay plate with three replicates of each. 5 µL RIPA + proteinase inhibitor was used as blank. 8820 µL BCA reagent A was mixed with 180 µL BCA reagent B to prepare the working reagent. 100 µL BCA working reagent was added to each well. The plate was covered in aluminum foil and incubated for 30 minutes at 37°C. The plate was brought back to room temperature, and the absorbance was measured in a plate reader at 562 nm. A standard curve was created by plotting blank-corrected standard absorbance versus protein concentration in µg/µL.

The standard curve was used to calculate sample protein concentrations, and then standardized by adding more RIPA buffer + proteinase inhibitor. 4x Laemmeli sample buffer was added to each sample to a final concentration of 1x under fume hood. The western blot was performed according to BioRad General Protocol for Western Blotting [98]. 59.0 µg of each sample protein was loaded into SDS page gel and run for 5 min at 50V and 150V for one hour until the samples were completely resolved. The gel was placed in transfer buffer for for 15 minutes. Transfer was performed according to the Trans-Blot Turbo Transfer System Transfer Pack[99]) (Bio-Rad, Cat#1704156) and Trans-Blot Turbo System (Bio-Rad). Roller was used to carefully remove any bubbles in the transfer sandwich. After transfer, the blot was incubated in EveryBlot Blocking Buffer (Bio-Rad, Cat.#12010020) for 30 minutes. Blocking buffer was removed and Primary antibodies Sigma ANTI-FLAG M2 antibody, Mouse monoclonal (Sigma, Cat.#F1804-50UG) was added to the membrane in the concentration 1:1000 antibody to EveryBlot Blocking Buffer. The membrane was incubated over night in 4°C on a platform rocker at slow gentle speed. Gel was put on a black tray and visualized in a Biorad ChemDoc Imaging System with manually adjusted exposure time between 10-300 seconds for the programs Chemiluminescent Blot and DyLight 650, DyLight 800, and IRDye 680RD.

The next day, the blot was washed three times with TBST (Sigma, Cat.#91414-100TAB) with 10 minute incubation in between each washing step. Secondary antibodies Goat Anti-Mouse Immunoglobulins/HRP (affinity isolated) (Agilent Technologies, #P0447), were added to blocking buffer in the concentration 1:10000. Blot was incubated for 30 minutes then washed again with TBST. Clarity Western ECL Substrate (Bio-Rad, Cat.#1705060) were mixed in 50 ml falcon tube 1:1. About 1 ml was used to develop the blot and incubated for 3-5 minutes.

Membrane were stripped by incubating the blot in Restore PLUS Western Blot

Stripping Buffer (Thermo Scientific, Cat.#46430) for approximately 10 minutes. Blot was then washed 3 times with TBST with 10 minutes incubation in between. The membrane was incubated in blocking buffer for 30 minutes. House keeping genes targeted with β-Tubulin Rabbit Polyclonal Antibody (LICORbio, Cat.#926-42211). The membrane was incubated over night in 4°C on a platform rocker at slow gentle speed. The next day, the blot was washed three times with TBST with 10 minute incubation in between each washing step. Secondary antibodies Goat Anti-Rabbit Immunoglobulins/HRP (affinity isolated) (Agilent Technologies, #P0448), were added to blocking buffer in the concentration 1:10000. Blot was incubated for 30 minutes then washed again with TBST. Clarity Western ECL Substrate (Bio-Rad, Cat.#1705060) were mixed in 50 ml falcon tube 1:1. About 1 ml was used to develop the blot and incubated for 3-5 minutes. Gel was was put on a black tray and visualized in a Biorad ChemDoc Imaging System with manually adjusted exposure time between 10-300 seconds for the programs Chemiluminescent Blot and DyLight 650, DyLight 800, and IRDye 680RD.

## 3.7 gRNA Secondary Structure Predictions

To predict secondary structures of the gRNA RTTs, the program RNAstructure[100] was used. The sequences of the template and homology arms only, were submitted to the portal and the most probable structure was noted along with the predicted free energy of the structure.

## 3.8 *In silico* Protein Engineering

This section will cover the rationale behind the protein engineering including engineering approaches for improving RT substrate affinity, and AI pipelines for thermostability and solubility improvements.

### 3.8.1 Rational Engineering

Rational engineering was employed to improve substrate affinity and structural integrity of the engineering subjects. To achieve this, three approaches were applied: transplantations, rational selection, and introduction of conserved residues (*figure 3.1*). The two conformations open-finger and closed-finger were studied for this purpose as they might pose different challenges for substrate affinity and binding during polymerization. Most mutations were introduced as single mutations unless there was enough reason to believe that a combination of certain mutations would be particularly beneficial.

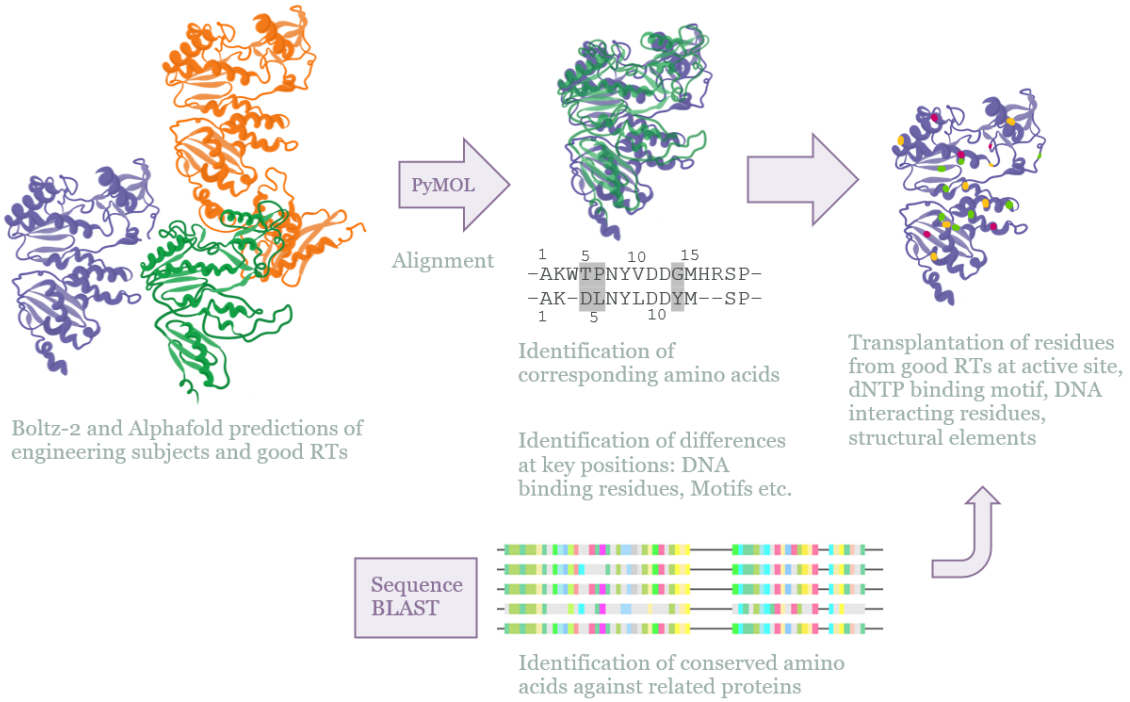



**Figure 3.1:** Rational engineering approach.

Transplantations of mutations and residues of interest were done to introduce elements from RTs with high TE efficiency to the engineering subjects. Both affinity enhancing mutations, and structure and flexibility mutations were of interest. To provide guidance for the engineering, MMLV PE2, MMLV PE6d, MMLV_MK, and Tf1 PE6c were used as references. All the RTs' closed-finger conformation structures were predicted using Boltz-2 with short complementary DNA and RNA sequences created from the MIF1 substrate used in the biochemical assays (MIF1_Boltz2_RNA and MIF1_Boltz2_DNA, sequences in *appendix A.1*). The prediction with the highest combined confidence score and ipTM score was selected for the analysis. These predicted structures were visualized in PyMOL. The engineering subject was aligned to either one of the reference protein structures, either by sequence aligning or superimposing the 3D structures. The position of an amino acid in the reference RT was translated to the corresponding amino acid in the engineering subject.

The structures were studied and compared to identify differences at key positions, for example DNA/RNA interacting residues and important functional motifs. For affinity mutations, the PyMOL integrated function to measure distances from amino acids to the substrate was used to highlight residues of interest within the range of interaction. Amino acids closer than about 7 Å to the substrate were studied in the references. Residues that were different: had stronger charge or different geometry, were transplanted to the engineering subjects to see how it affected affinity. Mutations from *table 2.1, 2.2 and 2.3* believed to affect affinity, were also transplanted to see if they would have the same positive effect on the engineering subjects.

Rational engineering were done based on fit, function, and local environment. Fo-

cus was put into changing (either increasing or decreasing) positive charge in the binding groove to increase or decrease affinity to the DNA backbone. Adjustments of amino acid lengths without changing the charge were also speculated to affect the affinity by bringing the DNA/RNA complex closer of further from the protein surface. Structural stability was improved by introducing mutations that would stabilize local environments such as helices, creating hydrogen bonds, and increasing hydrophobicity inside the protein.

Next, conserved residues from related proteins were identified. Related proteins from the PDB were sequence-aligned to the engineering subjects for comparisons. The most commonly conserved residues among these proteins that were not present in the engineering subject, were introduced.

Lastly, to study the open-finger conformation of the RTs it was necessary to construct a composite model of the proteins. To guide the creation of these models of the engineering subjects and references, a crystal structures of HIV-1 in finger-open-conformation[47] was used for superimposing the domains of the RTs in the correct conformation. Next, using the finger-open-composition of the engineering subjects, DNA/RNA complex and domain positioning was studied to find points of improvement, for example, replacing amino acids in risk of being steric hindrances, and improving flexibility using the reference RTs as a guide for transplantations.

### 3.8.2 Thermostability

The pipeline for introducing thermostability improving mutations included utilization of the AI tool ThermoMPNN. An overview of the process is outlined in *figure 3.2*. The predicted 3D structures and amino acid sequences of the engineering subjects were entered into ThermoMPNN, which put together a list of single and double mutations respectively, predicted to improve the thermostability of the protein. The list was sorted according to the mutations that would contribute with the largest improvement in free energy of the protein. Quality control was performed by checking the position of the best candidates in the 3D protein structure in PyMOL to avoid mutations that would disrupt important motifs or DNA/RNA binding. To diversify the selection, repeats of the same amino acid in single mutations, or repeats of the same mutation for double mutations were also avoided.

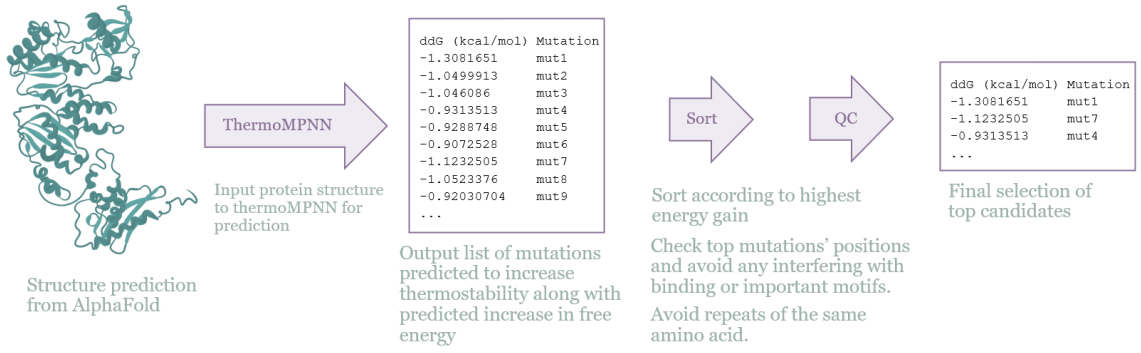



**Figure 3.2:** AI assisted thermostability engineering.

### 3.8.3 Solubility

To introduce solubility improvements in the RT structure, machine learning tool Soluble ProteinMPNN was used (*figure 3.3*). ProteinMPNN has previously successfully been used to improve properties such as stability and solubility of proteins [101], and was speculated to work on RTs as well. The first step of the pipeline included selecting residues to "lock", meaning residues that soluble ProteinMPNN would not change. To determine which amino acids to lock, DNA/RNA interacting residues, conserved residues and important motifs were studied. It was only desired to affect the surface residues of the protein, not disrupt any domains or interfere with function. Predicted protein structures from Boltz-2 with a short DNA/RNA complex of the MIF1 substrate used in the biochemical assays were used for the structure analysis (MIF1_Boltz2_RNA and MIF1_Boltz2_DNA, sequences in *appendix A.1*). Amino acids within 12 Å of the DNA/RNA according to the PyMOL software, was locked as to not interfere with the binding groove of the enzyme. DED motif was locked along with the active site and dNTP binding amino acids. Conserved regions were identified by blasting the RT sequence against NCBI reference genomic database. From the related sequences, a small random sample of 10 sequences was picked and aligned against the engineering subject's sequence. Highly conserved amino acids and regions were selected for locking.

After preparing the list with amino acids to conserve during the ProteinMPNN prediction, the list in *.jsonl* format together with a structural prediction of the protein from AlphaFold2 in *.yaml* format was submitted to Soluble Protein MPNN. The resulting list from the soluble ProteinMPNN prediction contained 300 sequences. The list was split into predictions with the highest sequence recovery and lowest sequence recovery to investigate whether low or high sequence recovery was more beneficial for increasing solubility while keeping enzymatic function. Then both of these lists were sorted according to global score, and 15 predictions with the highest global score from each of the lists were selected for further analysis. The top candidate sequences were predicted using AlphaFold2 and the confidence score of the predictions were noted. The predictive structure was also checked for misfolding. The top 10 candidates based on score and folding were selected for predictive analysis together with a DNA/RNA substrate using Boltz-2. These predictions were thoroughly studied. Structure alignments and comparisons to the original protein were made along with study of the vacuum electrostatic to make sure there were no hydrophobic patches or strongly charged surfaces. The confidence score along with ipTM score was noted. Taking all these criteria into consideration, three top candidates of each enzyme were selected.

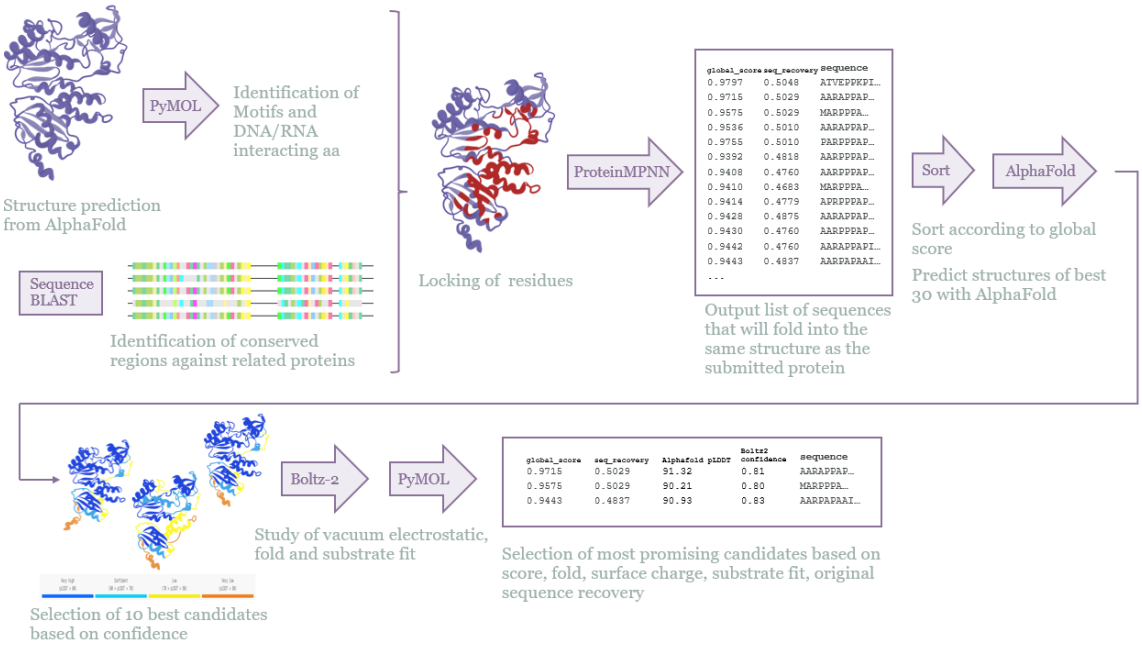



**Figure 3.3:** AI assisted solubility engineering.

# 4

# Results and Discussion

This chapter will cover and analyze the results of the characterization and benchmarking of the selected 18 RTs and then further go into the multivariable analysis of the data. Lastly, this chapter will cover the results of the rational engineering.

The structure of the project is outlined in *figure 4.1*. Throughout the project, two parts have been worked on in parallel. One part of the project includes screening of RTs for TE activity in HEK293T cells under different conditions. The benchmarking of the RTs required optimization of the functional reporter assay prior to starting. The next part includes *in vitro* biochemical and biophysical benchmarking of purified proteins. The information gathered through both these paths were used to inform the engineering of RTs. The engineering was done computationally through rational engineering and introduction of point mutations. Both rational engineering and AI tools were applied to carefully improve the desired properties. Before settling on a mutation, the variant was thoroughly studied and passed through multiple manual quality controls to increase the chances of success. Lastly, the engineered RTs were benchmarked in the established cellular assays for TE efficiency.

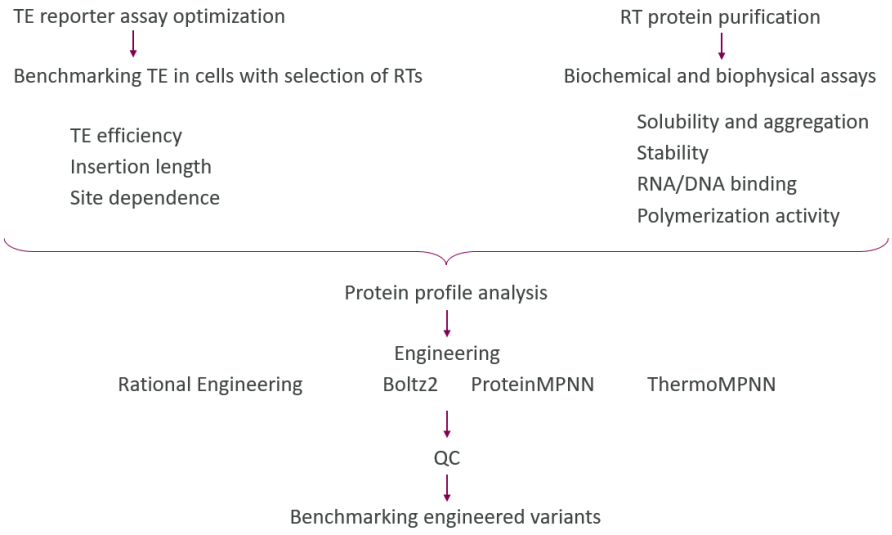



**Figure 4.1:** Overview of the project workflow combining complementary cellular assays with biochemical and biophysical analysis to gather data to enable informed rational engineering.

## 4.1 Reporter Assay Evaluation

For characterizing RTs, it would be beneficial to use a split TE system with the RT on a separate plasmid from the Cas9 component, as this system is more modular. The split editor would be more flexible and cost effective as only the RT-containing plasmid has to be switched while the SpCas9 plasmid is constant. To validate that the split editor resulted in satisfactory editing in comparison to the conventional tethered TE system, the tethered TE and split TE editor with identical RT (MMLV PE2) and Cas9 (SpCas9) components was compared on the same target site (*figure 4.2A*). The tethered editor had the PE2max architecture. The result showed that the split editor system had higher editing efficiency based on luminescence readout, than the tethered TE using the same springRNA. These results validated the use of the split editor for screening RTs.

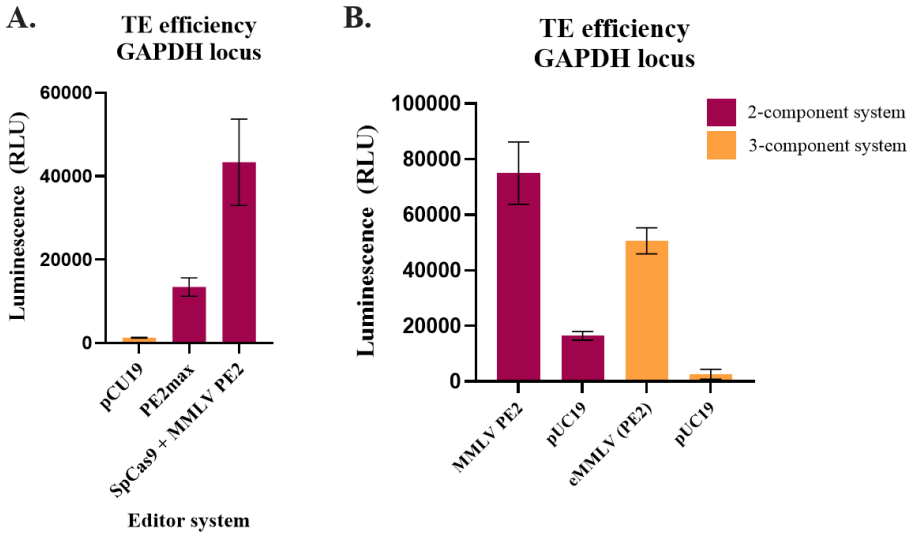



**Figure 4.2:** **A.** RT sequences are identical, and the tethered system has the PE-max architecture. RLU is not normalized to transfection efficiency as the PE2max plasmid did not have GFP expression. **B.** Comparison of 2-component system and 3-component system at the GAPDH locus.

Next, it was hypothesized that the sensitivity and workflow of the current HiBiT reporter assay could be improved by reducing the number of components, and selecting more efficient components for the system. The many components constructing the TE might reduce the transfection efficiency and also necessitate a more potent transfection reagent. The lysis step during readout also introduces a factor of unknown variability. Lastly, this method utilized a manual workflow which reduces the throughput of the characterization by the limitation of the use of 96 well plates. The small plates did not fit all RTs and replicates simultaneously and thereby prevented accurate, quantitative comparison of the TE efficiency. Addition of both the gRNA and LgBiT to one of the plasmids was investigated as a solution to to reduce the number of components and improve the workflow through endogenous expression of LgBiT circumventing the lysis step during readout. It was decided to

add these two components to the same plasmids as the SpCas9 to retain the full modularity of the system (*figure 4.3*). This 2-component system also allowed higher throughput by being compatible with a fully automated workflow from transfection to readout. Another advantage of this system was the switch to another transfection reagent. As the 3-component system used a synthetic gRNA, it required a more potent and thereby also slightly more toxic transfection reagent, however, removing the synthetic gRNA allowed the use of a more gentle transfection reagent.

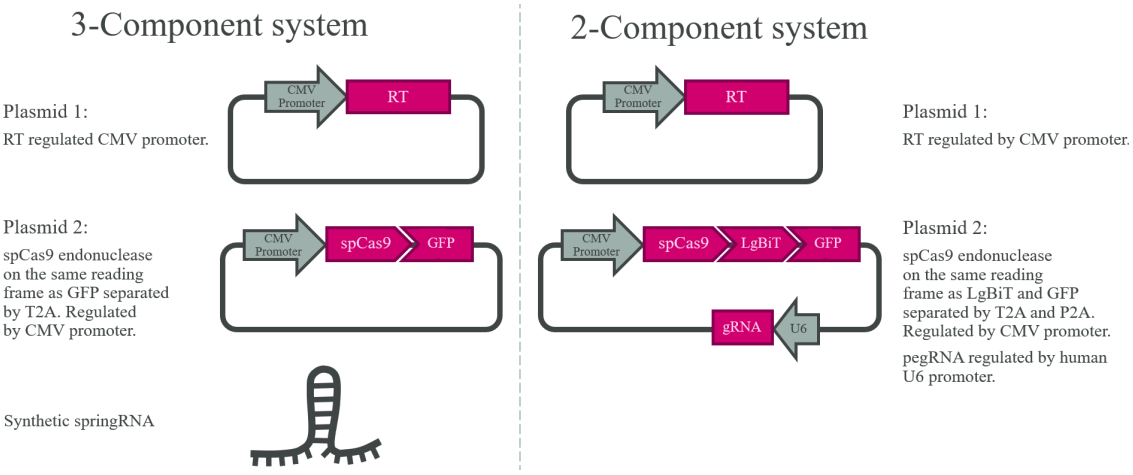



**Figure 4.3:** Comparison of 2-component system and 3-component system at the GAPDH locus.

The two systems were compared at the same target locus without compensation for their differences to decide if the new 2-component system would be as sensitive as the established 3-component system. Both systems were tested with springRNA targeting the GAPDH locus (*figure 4.2B*). The 3-component system shows slightly lower editing efficiency than the 2-component system using the reporter system, however, the 2-component system also has higher baseline luminescence according to the pUC19 mock editors. A possible reason might be background luminescence from the high endogenous expression of LgBit. The interpretation is that the 2-component editor has comparable or slightly higher editing efficiency than the 2-component editor at the GAPDH locus.

It was decided that the 2-component system would be used for benchmarking and characterization of the RTs. At the same time the gRNA structure was switched to pegRNA with homology arms as this structure is more conventional. Thus, I proceeded with screening a selection of 16 novel RTs for TE activity, to select 5 novel RTs of interest to base my further analysis on. AZRT_1 and its engineered version (eAZRT_1) was selected, along with two RTs (AZRT_2 and AZRT_11) that showed relative high TE activity at one locus. To add more nuance to the analysis, AZRT_15 was also included to allow analysis with a highly inefficient RT. These RTs were benchmarked together with the panel of RTs at two target sites to compare editing efficiency and evaluate site specificity (*figure 4.4A*). To make sure that the RT would be the only determining variable of the editing outcome,

all components of the editor were expressed at very high concentrations using the strong CMV promoter.

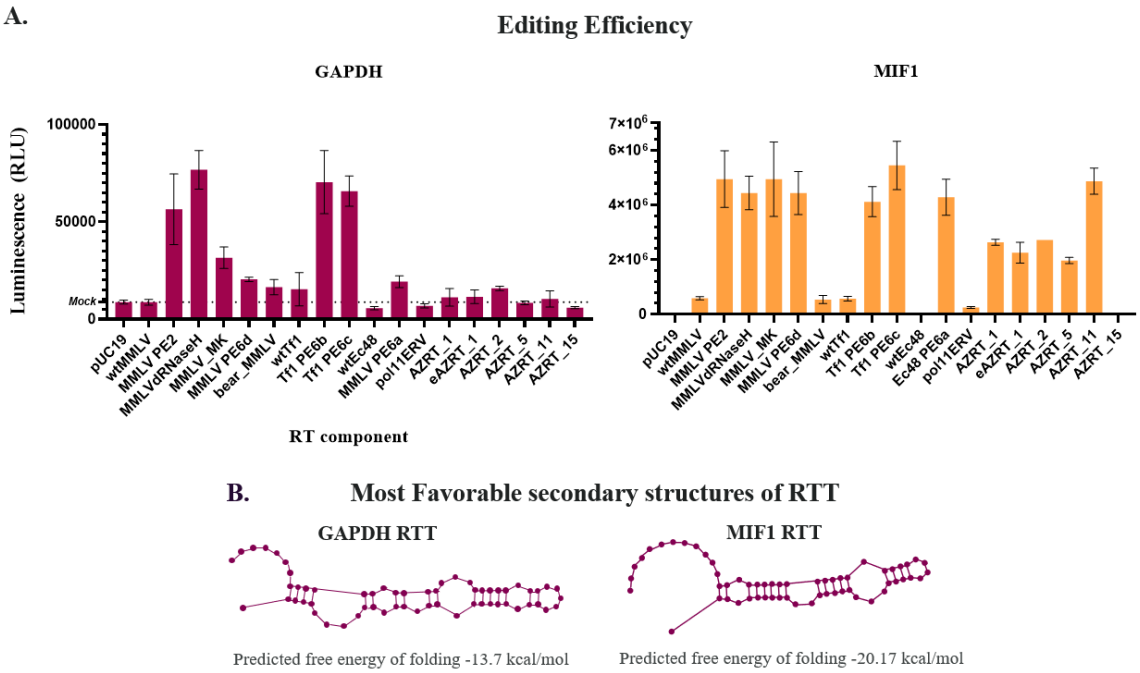



**Figure 4.4:** **A.** Editing efficiency with saturated TE system. (SEM, n=1 to 4). **B.** Predicted secondary structures of the gRNA RTTs for GAPDH and MIF1 loci generated with RNAstructure[100].

The data in *figure 4.4A* implies that many of the RTs are highly site dependent. At the GAPDH site, the RTs MMLV PE2, MMLVΔRNaseH, Tf1 Pe6b and PE6c have high editing efficiency in the reporter assay. These four RTs also exhibit high editing efficiency at the MIF1 site. On the other hand, the AZ RTs, and engineered MMLV PE6d, Ec48 PE6a, and MMLV_MK show low editing efficiency on the GAPDH site. Some of these RTs does however perform very well on the MIF1 target site, including: PE6d, PE6a, and MMLV_MK, which show comparable editing efficiency to the MMLV PE2 at the MIF1 locus. AZRT_11 also show a significant increase in editing efficiency at the MIF1 site, reaching levels comparable to the MMLV PE2. Due to the antisense strand targeting of the MIF1 gene, it is possible that RTs that perform scaffold incorporation have reduced reported efficiency at the MIF1 site. However, it seems that most RTs have higher efficiency at the MIF1 locus, thus, directly disagreeing with that statement.

To investigate the reason for the site dependency, secondary structures of the RTTs (homology arms and template) of the gRNAs were investigated following the approach by Doman J.L et al.[11] (*figure 4.4B*). Previous research has proposed that the level of processivity of the RT might be important for TE efficiency. Doman J.L. et al. proposed a theory of the correlation between RT processivity and RTT secondary structure, suggesting that RTs with high levels of processivity were more

robust on structured RTTs while unprocessive RTs increased the generation of prematurely terminated, unproductive, RT products when faced with a highly structured RTT substrate [11]. On unstructured edits, however, more processive RTs increase the risk for indels and scaffold incorporation. Without directly assessing processivity, RTs MMLV PE2, Tf1 PE6c, and MMLV PE6d were categorized as having a high level of processivity, while RTs Tf1 PE6b and MMLVΔRNaseH were said to have lower processivity. A guideline was established to determine when a processive RT would perform better than one with lower processivity by looking at the free energy of folding of the RTT. If the energy was greater than -23 kcal/-mol, a more processive RT performs significantly better, but if it is lower than -23 kcal/mol a less processive enzyme might enhance editing outcomes [11]. The theory was only tested on MMLVΔRNaseH and MMLV PE6d however, and might not be representative for all RTs.

The RTT of the MIF1 site (*figure 4.4B*) would be relatively structured according to my interpretation of Doman J.L. et al.'s definition, with the most favorable secondary structure having a free energy of folding of -20.17 kcal/mol. The GAPDH RTT also showcase a tendency to form secondary structures, but is in comparison to the MIF1 template more unstructured with a predicted free energy of folding of -13.7 kcal/mol. Comparing MMLV PE6d and MMLVΔRNaseH, MMLVΔRNaseH should perform better than MMLV PE6d on the unstructured template according to Doman J.L.'s tests, which agreed with the obtained data. However, according to the same tests, MMLVΔRNaseH would have more trouble than MMLV PE6d on the MIF1 RTT. Which does not seem to be the case. Though, the MIF1 RTT does not reach the guideline value established by Doman J.L. et al. meaning that the hypothesis might still be true.

Another factor that might influence the editing efficiency is the kinetics of the different target sites. It is possible that cutting efficiency is different at the two target sites, or that the sites does not reach maximum editing at the same time. Another potential reason would be the different expression levels of the target genes. MIF1 is expressed at about 50% higher levels than GAPDH which means that the MIF1 locus would be more sensitive to report lower editing efficiencies, as even low editing would result in high expression of the HiBiT tag.

Another hypothesis is that the editing efficiency is affected by the concentration of the RT. The disadvantage of overexpression of the TE components when the properties and behavior of the RTs are unknown is that high concentrations of the RT might have unknown effects. I speculate that overexpression of some of the less soluble or aggregation prone RTs might cause dimerization or aggregates, reducing the concentration of functional RTs in the cells, effectively leading to reduced editing efficiency.

## 4.2 RT Dose-Response

To test the hypothesis regarding whether the RT concentration affects TE activity, plasmids with the RT under 9 different promoters were designed in addition to the CMV promoter to do endogenous titrations of the RT component in cells. First, this was tested on the MMLV PE2 RT to confirm that the promoters yielded a dose-response curve for TE efficiency in cells (*figure 4.5A*). The results show a dose response, under identical conditions, the editing efficiency in the reporter assay varies greatly depending on the promoter used to express the RT on both target sites. These results indicate that RT concentration does indeed affect editing efficiency.

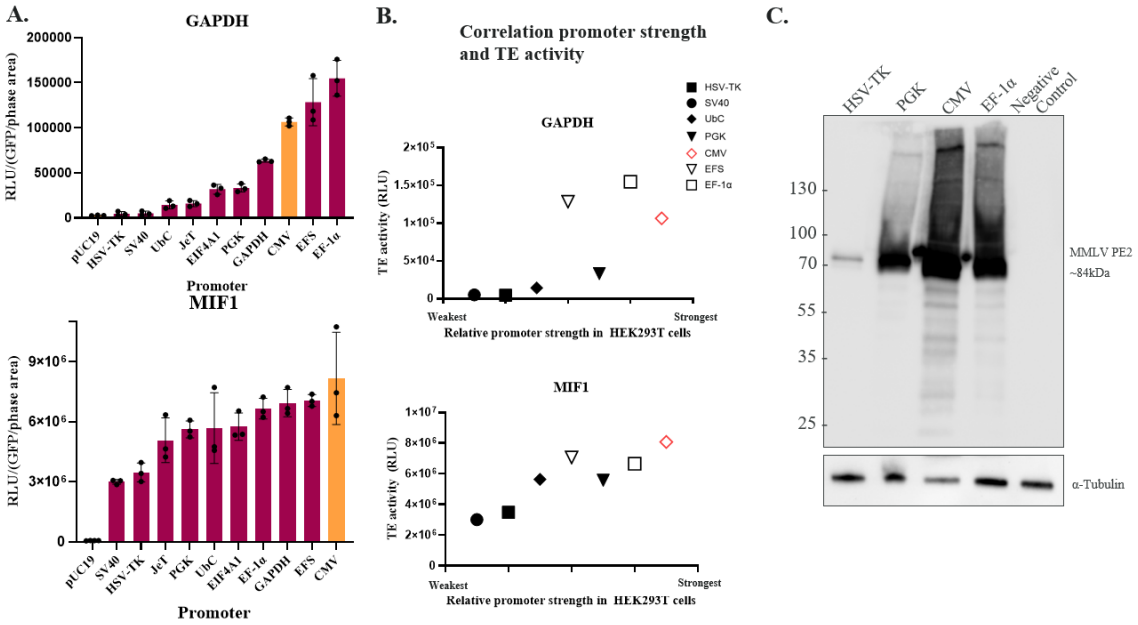



**Figure 4.5:** **A.** Titration of MMLV PE2 in cells at the two target loci GAPDH and MIF1. (SD, n=3) The GAPDH site was tested with the 3-component system while the MIF1 site was tested with the 2-component system. **B.** Correlation between TE efficiency and supposed promoter strength in HEK293T cells [73]. **C.** Western blot of expression of MMLV PE2 under the regulation of the promoters HSV-TK, PGK, CMV, and EF-1$\alpha$ extracted from HEK293T cells 72h post-transfection.

The results of the endogenous RT titration agrees with the anticipated effect: the stronger promoters generally elicit higher editing efficiencies than the weaker promoters (*figure 4.5B*). However, the strongest promoter does not always yield the highest editing efficiency. Interestingly, it seems like the effect of RT expression is different at the two target sites. At the GAPDH locus, a maximum editing plateau is never reached. It seems as if 72h is insufficient for obtaining maximum editing efficiency at this site. At the MIF1 site, however, editing might be much more efficient, reaching higher TE efficiency with lower RT expression, possibly due to the gene's high expression levels in comparison to GAPDH. There might even be a plateu forming with the EF-1$\alpha$, GAPDH, EFS, and CMV promoter, indicating a maximum editing efficiency for MMLV PE2, however, the standard deviation of the

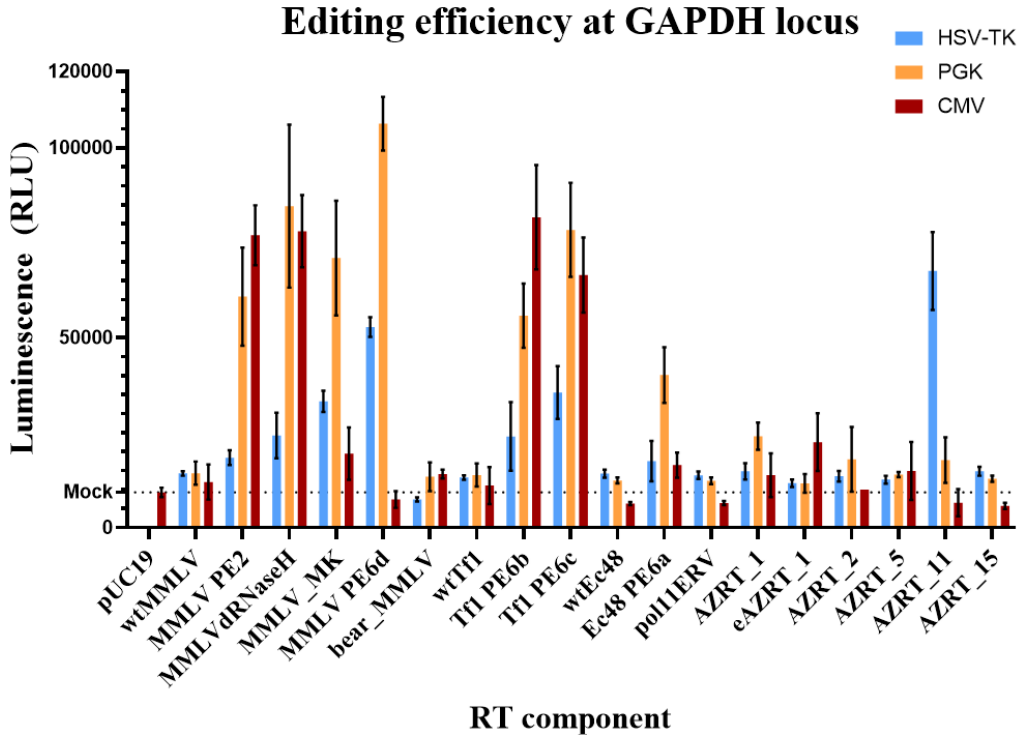
CMV promoter is too wide to draw any conclusions. One obvious possibility that could answer for this difference is that the sites were tested with different editor systems, the 3-component vs the 2-component systems, and the systems might have different editing kinetics. In this case comparisons between the target sites might be futile.

Another possibility for the differences, is that the kinetic of the promoters have different influence at the different target sites. At the GAPDH site there is no correlation between promoter strength and TE activity, while the data from the MIF1 locus indicates a positive, linear correlation. To confirm the actual expression of MMLV PE2 in the cells at 72h the MMLV PE2 protein was extracted and blotted to compare the protein levels in the cells depending on the four promoters CMV, EF-1$\alpha$ and HSV-TK (*figure 4.5C*). The results agree with literature. The results shows a higher concentration of MMLV PE2 with the CMV promoter, followed by EF-1$\alpha$, PGK, and last, HSV-TK, which gives very weak expression of the enzyme in comparison to the other promoters. These results indicate that the strongest promoter might not ensure the highest level of editing, suggesting issues related to the RT such as aggregates during overexpression reducing editing efficiency, or that the kinetics of the promoters might have influence over editing efficiencies.

At the GAPDH locus it seems the highest editing efficiency is obtained with promoters that are weaker than CMV: EFS and EF-1$\alpha$. Both of these promoters have reported stable high expression over long periods of time, but also reaches maximum expression relatively quickly in contrast to the CMV promoter that has immediate strong expression that is less stable over time. PGK which also has stable long-term expression, enacts lower editing efficiency, possibly due to lower initial expression. This suggests that editing at the GAPDH site might be favored by strong initial expression, which would mean that overexpression is constructive to editing with MMLV PE2. Another advantage of not overexpressing the RTs when characterizing their TE efficiency, is increased resolution of the assay, which would allow more accurate comparisons of the RTs.

To investigate whether other RTs were also potentially negatively effected by overexpression, three promoters of varying strengths were selected: CMV, PGK, and HSV-TK. And the dose response was tested in HEK293T cells (*figure 4.6*). The results indicate that there are some RTs that performs better with weaker promoters with lower initial expression. One notable case is MMLV PE6d which performs better with the weaker promoters on both target sites, indicating an potential issue with overexpression of the RT. Possibly, the RT might be unstable and aggregate at high concentrations due to misfolding. There are also some other RTs which exhibits higher editing efficiencies with the weakest promoter, such as AZRT_11 at the GAPDH site. AZRT_11 does however perform similarly with the PGK and CMV promoter at the MIF1 target site. There are also a few RTs that perform significantly better with the stronger promoter, such as Tf1 PE6c, Ec48 PE6a and both AZRT_1s, but only when editing the MIF1 gene. On the GAPDH locus, editing efficiency is more similar between the stronger promoters. An advantage of

screening the RTs with a weaker promoter is that it might improve the resolution of the HiBiT assay. When the system is saturated, nuances between RTs might disappear as the quantity of the RT compensates for the lack of efficiency letting meaning all RTs might reach the "roof" of highest possible luminescence even though their efficiencies are very different. Due to convenience and construct availability, the CMV promoter was still used in the remaining screening.



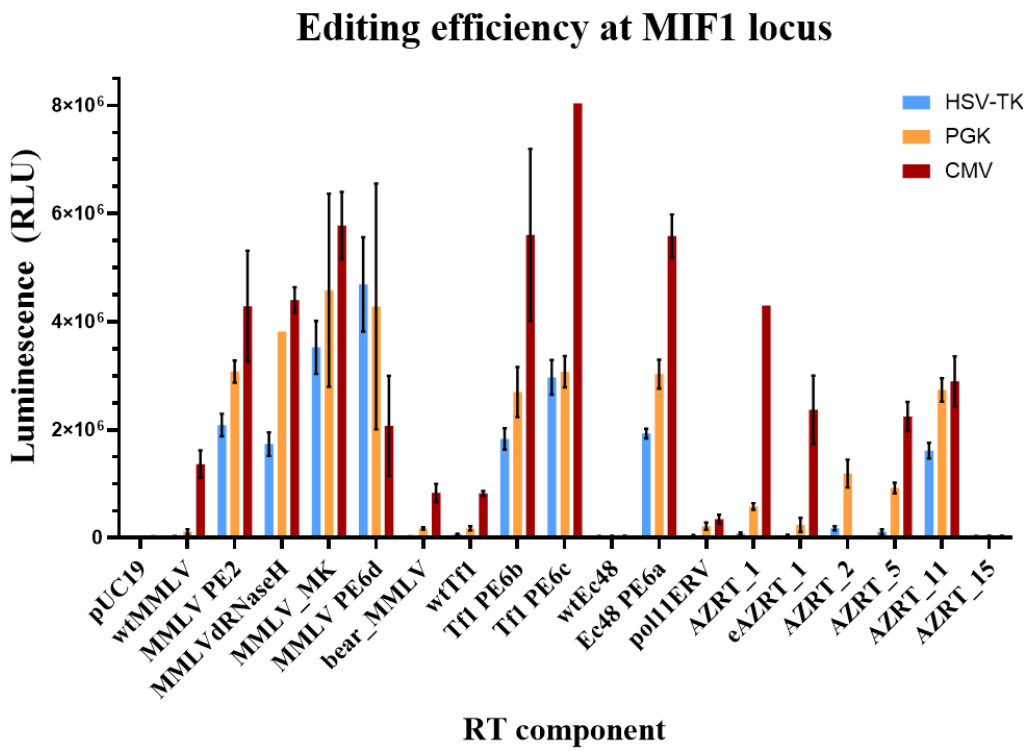



**Figure 4.6:** TE efficiency based on RT expression using promoters HSV-TK, PGK and CMV. (SD, n=1 to 4)

## 4.3 Insertion Length

To test the RTs' capacity for installing longer templates, a new gRNA for the 2-component system was designed. The template contained a filler sequence (3x FLAG) 3' prime of the HiBiT tag (*figure 4.7A*). The rationale was that the RT would have to successfully insert the full RTT for HiBiT to be expressed, which meant that only complete insertions would be detectable in the reporter assay. The longer gRNA's effect on TE activity was investigated at the MIF1 locus using CMV promoter to express the RTs. Inserting longer templates reduced editing efficiency significantly for MMLV PE2 (*see figure 4.7B*).

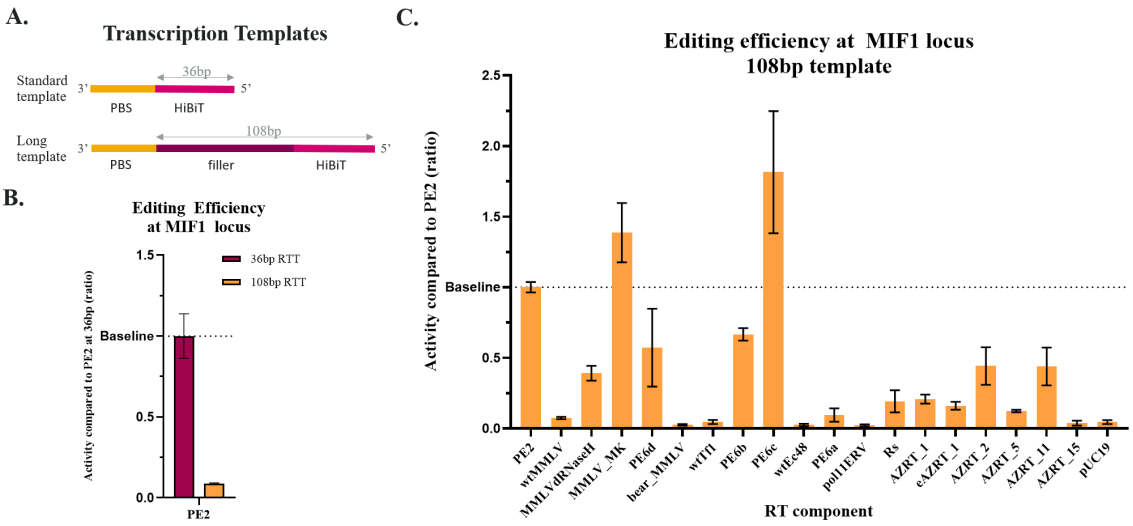



**Figure 4.7: A.** Template designs. 36bps (standard HiBiT insertion) and 108bps (filler + standard HiBiT). **B.** Comparison of the editing efficiency of MMLV PE2 at MIF1 locus using two different length RTTs. Bars show standard error of the mean for n=3 biological replicates. **C.** TE activity with 108bps RTT for RTs normalized to activity of PE2. (SD, N=3)

*Figure 4.7C* shows how well the RTs insert the longer template in comparison to MMLV PE2. The data indicates that Tf1 PE6c and MMLV_MK handles longer RTTs significantly better than MMLV PE2. These results agree with the expectations as Tf1 PE6c previously been reported to be more efficient than MMLV PE2 at extending long templates [11]. Since both Tf1 PE6c and MMLV_MK possesses mutations to improve substrate affinity and potentially processivity, the results in *figure 4.7C* indicates that the capability to reverse transcribe longer templates might be related to these properties. MMLVΔRNaseH and Tf1 PE6b, that presumably has a lower levels of processivity than MMLV PE2, performs accordingly in this assay, with lower editing efficiency.

## 4.4 Protein Purification

The RTs were extracted from *E. coli* overexpression cultures and purified through multiple steps to obtain pure proteins for *in vitro* assays. A SDS PAGE of the supernatant after lysis of the *E. coli* is shown in *figure 4.8A*. Soluble proteins will be present in this phase and can then be purified through multiple purification steps.

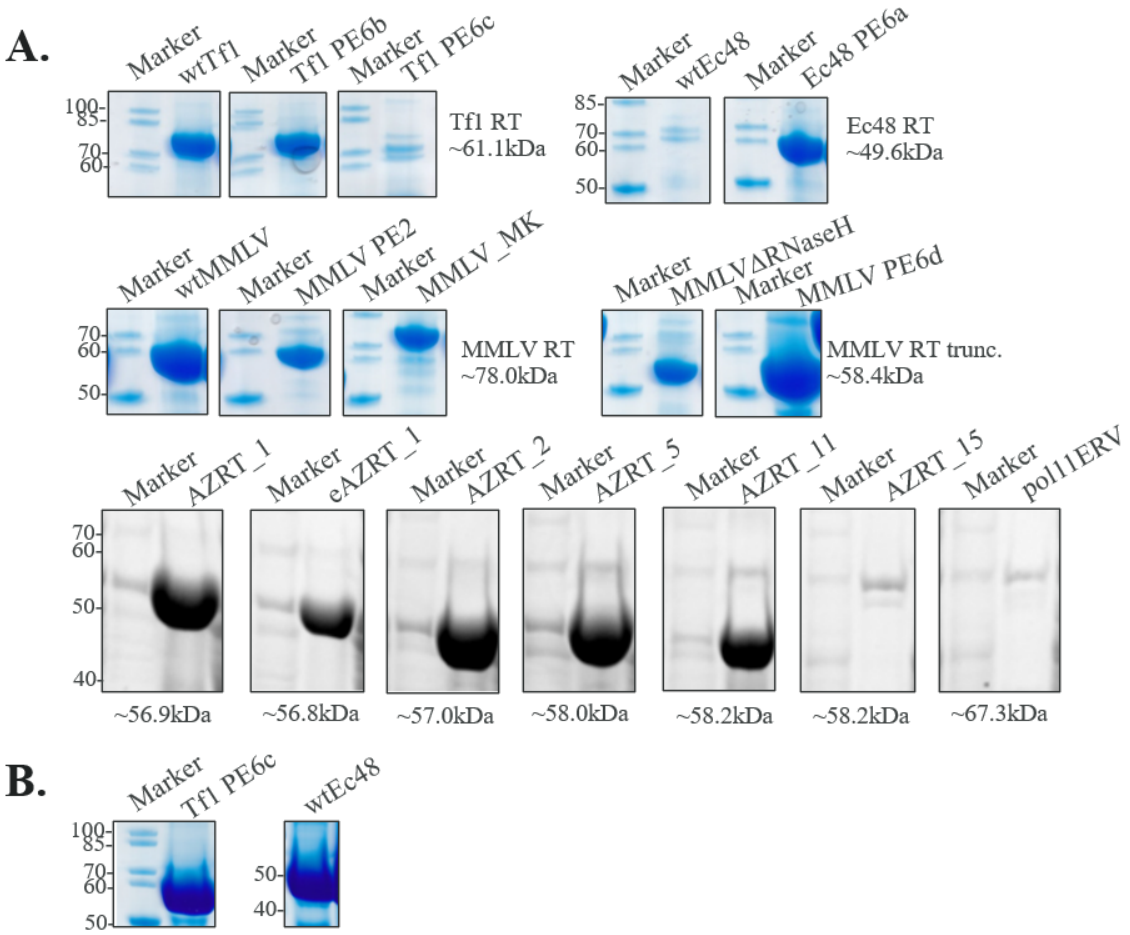



**Figure 4.8: A.** SDS PAGE of cellular proteins extracted from *E.coli* after recombinant RT expression. **B.** Repeat after extraction from exclusion bodies.

The SDS Page shows that most RTs were present in the supernatant indicated by the bands on the gel (*figure 4.8A*). However, Tf1 PE6c, wtEc48, AZRT_15 and pol11ERV are missing due to formation of inclusion bodies, suggesting misfolding and intrinsic solubility issues. Further supporting this hypothesis is that Ec48 originates from *E. coli*, and should thus not misfold when expressed in *E.coli*. This indicates that it probably has a structure prone to forming aggregates even in it's native environment. Tf1 PE6c and wtEc48 were successfully extracted from the inclusion bodies (*figure 4.8B*) and further purified. No repeated extraction of AZRT_15 and pol11ERV was done due to time constraints, and thus the RTs could not be characterized in *in vitro* assays. After successful extraction and purification using size exclusion chromatography (SEC), the purity was checked using SDS PAGE. The purity of each RT was confirmed to be adequate for the following biochemical and biophysical assays for all RTs except Ec48 PE6a. Due to the time constraints purification could not be attempted again for this RT. The potentially impure Ec48 PE6a was still investigated through *in vitro* assays, but the results might not be representative for the RT. All RTs were stored in the same buffers due to not knowing optimal conditions for the RTs.

AZRT_1, eAZRT_1, and AZRT_2 were extracted at very high concentrations which was evident early from the high viscosity of the extract, and later confirmed through concentration measurements. All three protein solutions became cloudy and formed aggregates during purification.

## 4.5 Aggregation

Due to suggested solubility issues, I investigated aggregation tendencies of the RTs through turbidity measurements during heating (*figure B.1*). As the protein denature they will form aggregates which turns the solution cloudy. However, buffer conditions will also affect stability of the enzymes which might have impacted the RTs differently depending on their optimal conditions. The measurements of Ec48 PE6a and AZRT_5 are possibly not representative due to unreliable measurements of turbidity (*figure B.1A*).

The aggregation temperatures of the RTs range from around 34 to 52°C (excluding AZRT_5 as an error) (*figure 4.9*). It seems the Tf1 family has a tendency to aggregate at low temperatures compared to the other RT families and AZ RTs, falling under the physiological temperature of biological systems. wtEc48 and Tf1 PE6c that both formed inclusion bodies, differ a few degrees in aggregation temperature (34 and 37 respectively), while wtTf1 and Tf1 PE6b which have even lower aggregation temperature and no issues with solubility in *E.coli*. Thus, failing to directly explain the forming of inclusion bodies in *E.coli*. The low aggregation temperature for wtTf1 and Tf1 PE6b might also be related to poor stability or suboptimal buffer conditions causing the protein to aggregate *in vitro*. AZRTs also tend to form aggregates at relatively low temperatures (*figure 4.9*). The engineered RTs of the MMLV family seem to be very resistant to aggregation, especially the bear_MMLV which is exceptionally stable.

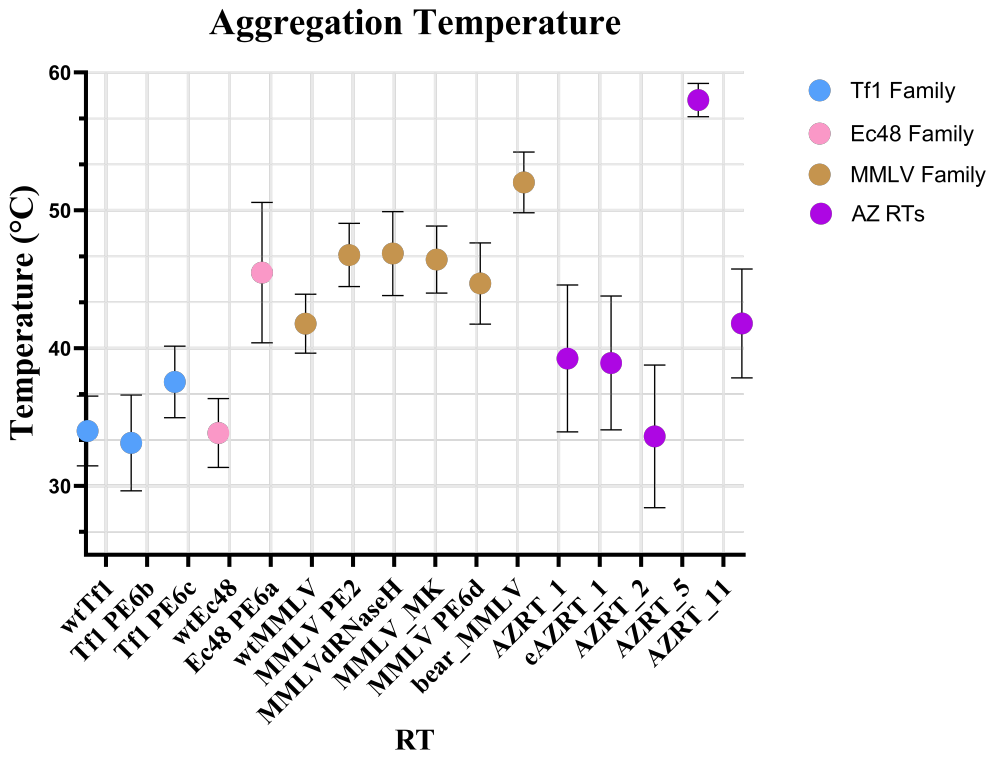



**Figure 4.9:** Aggregation temperature of purified RTs. Measured through turbidity measurements during heating. (SD, n=4)

## 4.6 Thermostability

To assess the stability of the purified RTs, thermostability was measured using thermal shift assay, and tryptophan fluorescence (*appendix B.2*). The two methods did not agree about the specific melting points (*table B.1*), however, this might be due to different buffer conditions used for the two methods. When measuring tryptophan fluorescence the concentration of the surfactant Tx100 solubilize the proteins and prevent protein-protein interactions meaning that stability of the proteins might be affected. The two methods did however agree on the general trends among RT families. With tryptophan fluorescence there were a few proteins where the measurements were highly unreliable due to noise. In the end, thermal shift assay was determined the most reliable and representative as it is less affected by buffer conditions. However, this method was not suited to measure AZRT_5 due to the low concentration of the stock of the protein.

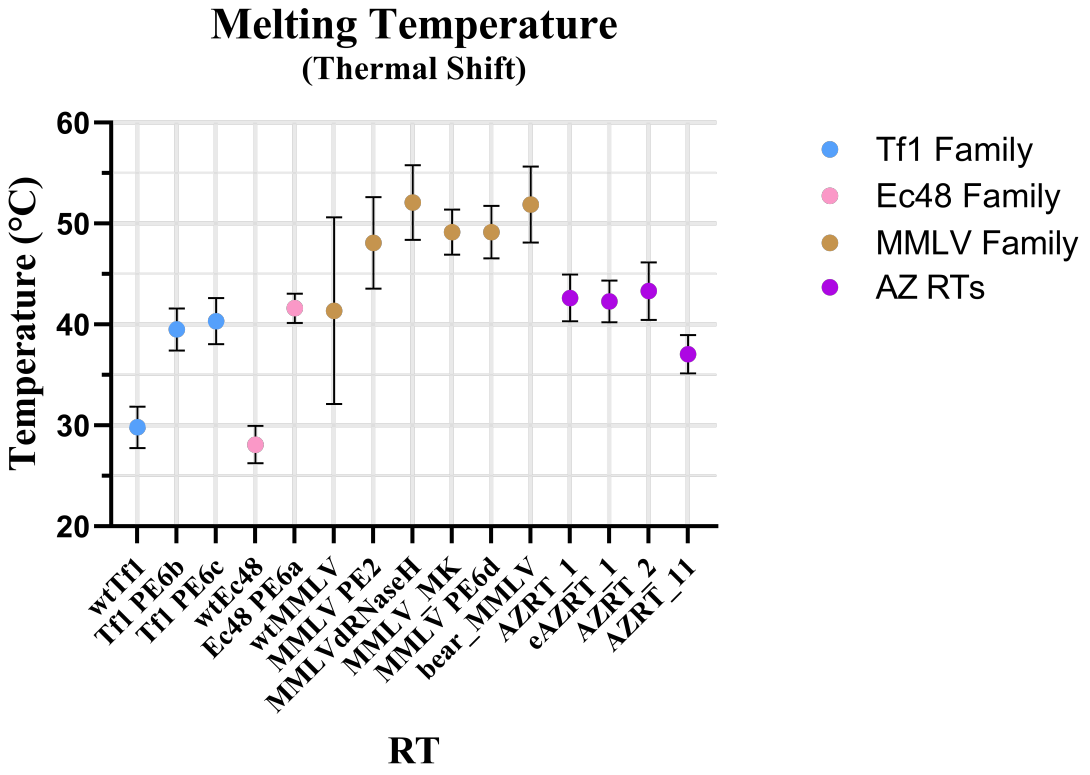



**Figure 4.10:** Melting temperature (Tm) of purified RTs. (SD, n=3)

The thermal shift assay resulted in regular melting curves (*figure B.2A-D*) that implied that the least thermostable RTs were wtTf1 and wtEc48, both unfolding already under 30°C while the most thermostable proteins were both MMLV variants with Tm around 52°C (*figure 4.10*). The melting points indicate an increase in stability of the engineered variants compared to wild type RTs. wtMMLV have large standard deviation which comes from a difficulty in approximating the true melting point from the melting curves. Both wtMMLV and MMLV PE2 might have two melting events taking place (*figure B.2C*). While bear_MMLV and MMLVΔRNaseH share the highest thermostability, bear_MMLV have low TE efficiency while MMLVΔRNaseH have high efficiency, demonstrating that high stability does not necessarily equal high TE efficiency. Interestingly, the wtMMLV and engineered PE6a, PE6b, and PE6c (all efficient editors on at least one target site according to data in *figure 4.4A*) have similar melting points, further indicating that the level of stability does not determine whether the RT will be effective as part of a TE. All engineered RTs do however, have greater stability than their wild type counterparts, suggesting that improving thermostability seems to have an effect on editing efficiency.

## 4.7 DNA/RNA Interaction

To measure DNA/RNA binding, a fluorescence polarization assay was established with a FAM-labeled substrate, which measures the change in light polarization when the RT binds the substrate. The goal of the assay was to obtain the equilibrium dissociation constant Kd, meaning the concentration of protein necessary to bind half of the substrate at equilibrium. First, key variables had to be optimized, such as $MgCl_2$ concentration, substrate concentration, salt concentration, and protein concentration interval (*figure B.4*). For $MgCl_2$, it was determined that binding was the most favored at a concentration of 1 mM (*figure B.4B*), which also is within the range of free cytosolic $Mg^{2+}$ available in mammalian cells [102]. To increase sensitivity for lower concentrations of protein, a low amount of substrate is desirable. If the substrate concentration is lowered, it would enable the capturing of the entire binding curve using a lower protein concentration, which is important as the concentrations of the protein stocks are a limiting factor. To obtain the full binding curve, including the plateau where more protein does not increase the binding interaction, it was necessary to use as high concentrations as possible for each RT. However, whether measurements with lower concentrations of substrates are possible, depends on the combination of fluorophore and instrument sensitivity. For this experiment, it was possible to use 1 nM substrate for the measurements and still obtain reliable binding curves (*figure B.4C*). Many concentration intervals were tested for the binding curves.

The DNA/RNA substrates designed for assessing binding, were shorter stretches of the MIF1 and GAPDH genes. The templates were designed to mimic the genes and did not consider stability. The DNA primer and bases that annealed to the RNA template, had the identical sequence to the PBS annealing to the genomic DNA in the cells. Multiple substrate variants were tested until one representative of each locus was selected. These different variants included different lengths, overhangs, and Y-shaped substrates. In the end, it was decided to use substrates without overhang for benchmarking of binding interactions. The substrates for the two genes were not identical in their design, and had both different template lengths, and primer lengths (*figure 4.11A*).

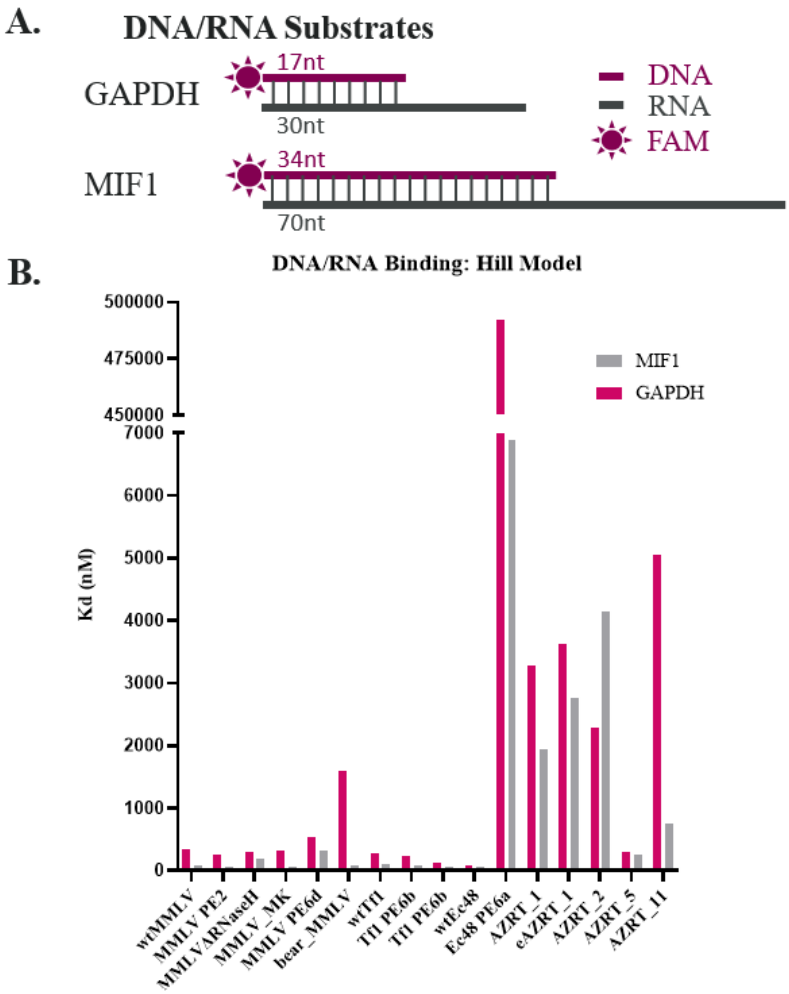



**Figure 4.11: A.** DNA/RNA substrates used for benchmarking in the fluorescence polarization assay. **B.** Kd for RTs binding to GAPDH and MIF1 DNA/RNA substrates.

RTs were benchmarked for binding interaction with GAPDH and MIF1 substrates and the resulting data was fitted to the Hill model and the Two sites binding model to find the model that described the data best (*table B.2* and *B.3*). It was determined that the Hill model fitted the majority of RTs and was therefore used to obtain Kd (*figure 4.11B*). The binding data suggests that the RTs bind less strongly to the GAPDH DNA/RNA substrate than the MIF1 substrate, indicated by higher Kds when interacting with the GAPDH substrate. The only exception being AZRT_2. Next, there are a few RTs that have higher Kds than the other RTs in general, these are the bear_MMLV, Ec48 PE6a and all the AZ RTs except AZRT_5. Ec48 PE6a is suspected to not be an accurate representation of the enzyme's actual properties as an effect of possible impurities in the protein sample. Ec48 PE6a never reached the plateau of maximum bound protein during binding regardless of protein concentration (*figure 4.12* and *4.13*). In contrast, all wtRTs possess very low Kds for both substrates, most are comparable to their engineered variants or even lower.

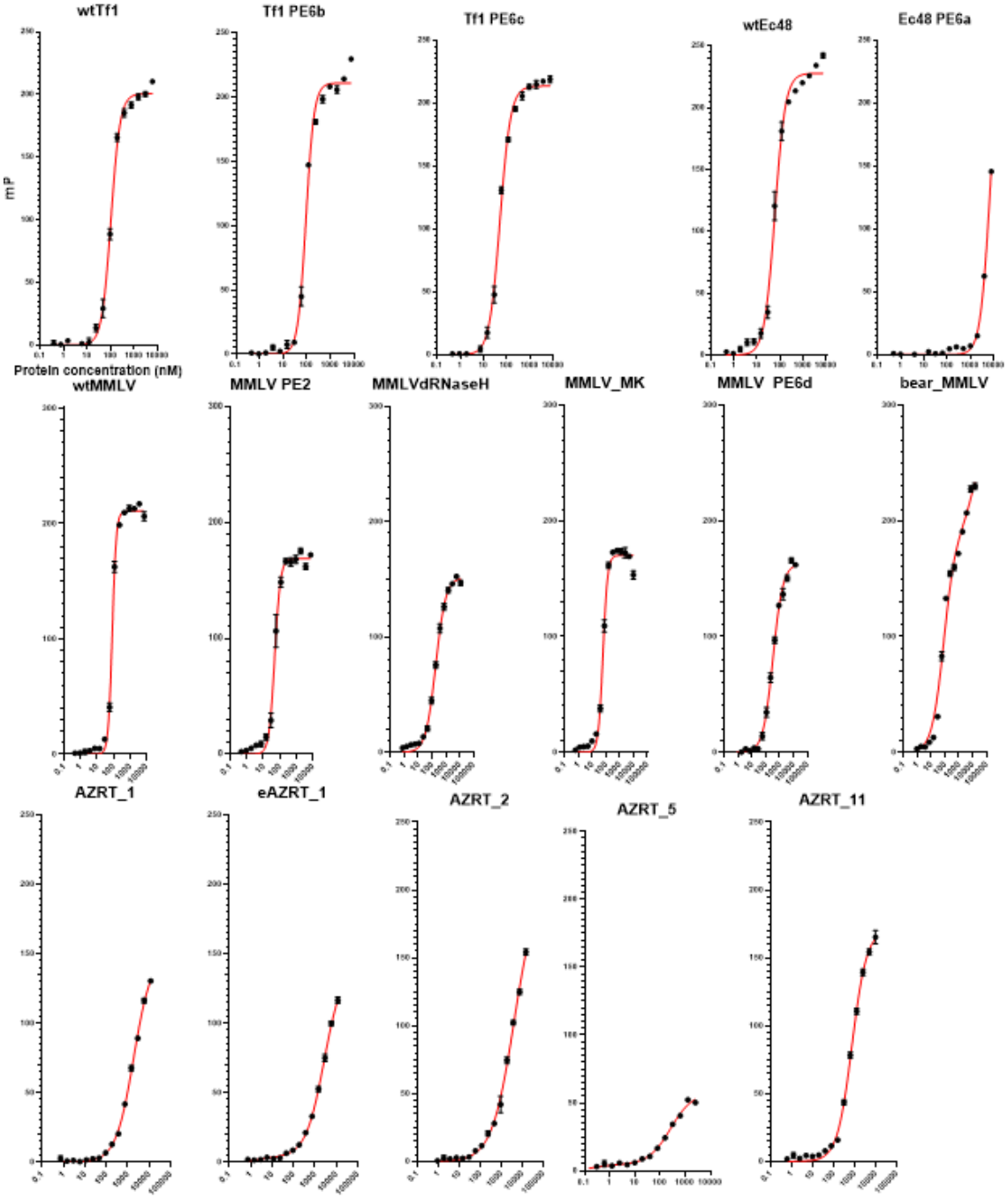



**Figure 4.12:** Most probable model fitted to fluorescence polarization data on MIF1 substrate.

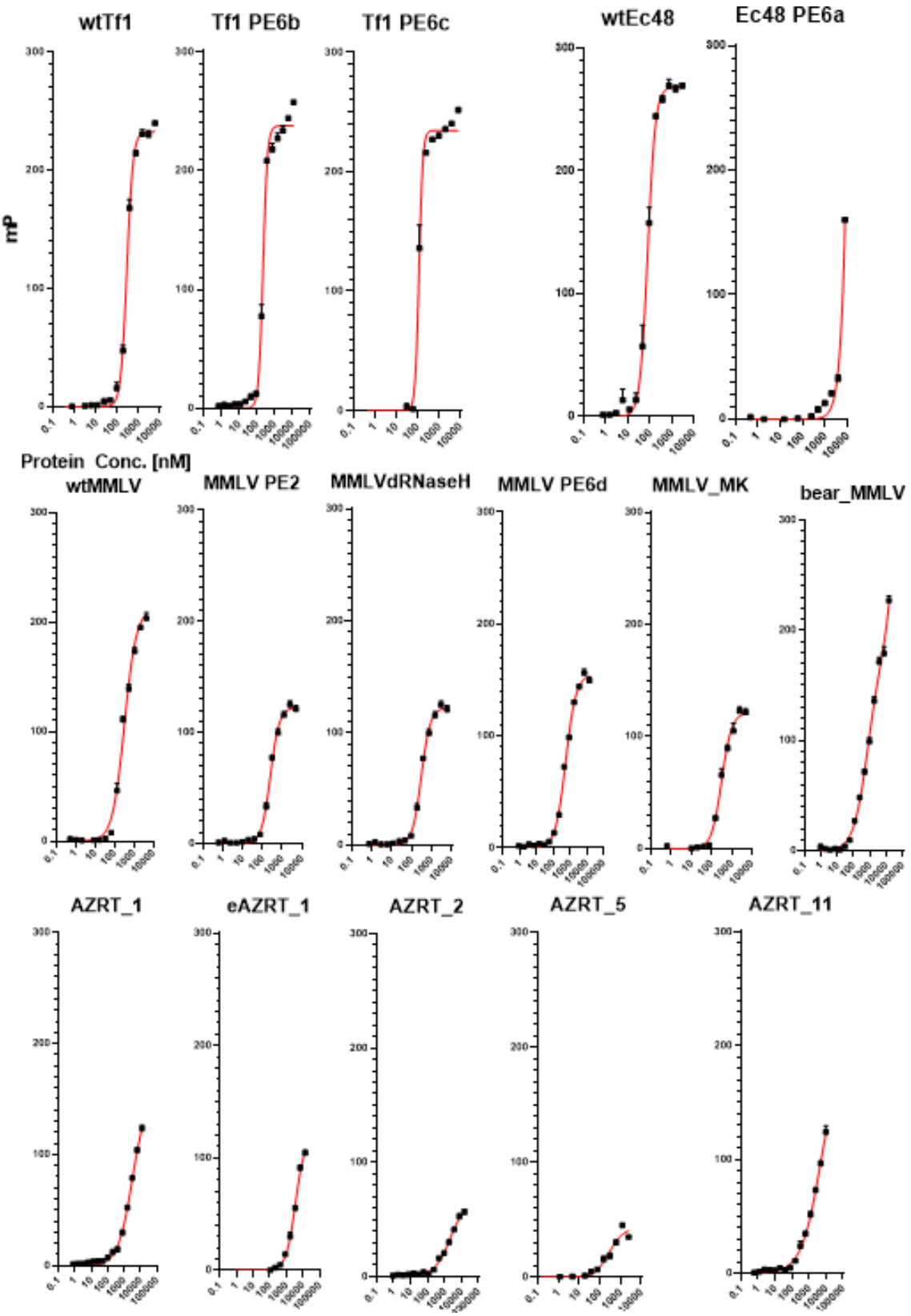



**Figure 4.13:** Most probable model fitted to fluorescence polarization data on GAPDH substrate.

After assessing DNA/RNA binding, affinity to other nucleic acids was investigated. The aim was to investigate whether DNA/RNA complexes were the preferred substrate. Therefore the fluorescence polarization assay was repeated with DNA/RNA of the MIF1 gene. This time, binding interaction with ssDNA, dsDNA, ssRNA, and dsRNA of the same sequence as the DNA/RNA substrate was also tested. The resulting benchmarking data showed that all RTs have a surprisingly similar affinity to any nucleic acids (*figure 4.14*), which means that the RTs in the cells not exclusively bind to the desired hybridized PBS to initiate polymerization. In general, it seems that most RTs have the weakest affinity to dsRNA, while they bind the strongest to ssRNA. For TE, this might translate into two scenarios, either the RT binds the gRNA attached to the SpCas9, or it might bind mRNA in the nucleus with higher affinity than the intended DNA/RNA complex. It is hard to guess what effect this actually has in the cells during TE, however, it might be interesting to look into reducing the affinity to other genetic material in an attempt to improve TE activity. To further analyze the nucleic acid binding, random sequences of nucleic acids should be used to eliminate sequence specificity as a factor.

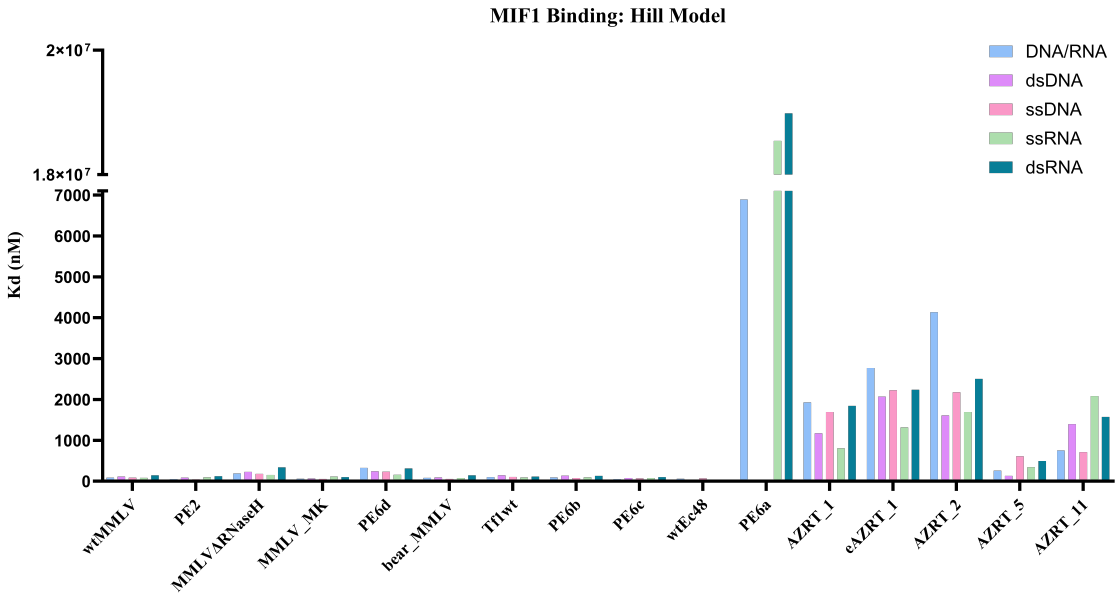



**Figure 4.14:** Binding interaction between RTs and different nucleic acids. Data points are missing for Ec48 PE6a dsDNA and ssDNA due to inability to fit the FP data to the Hill Model.

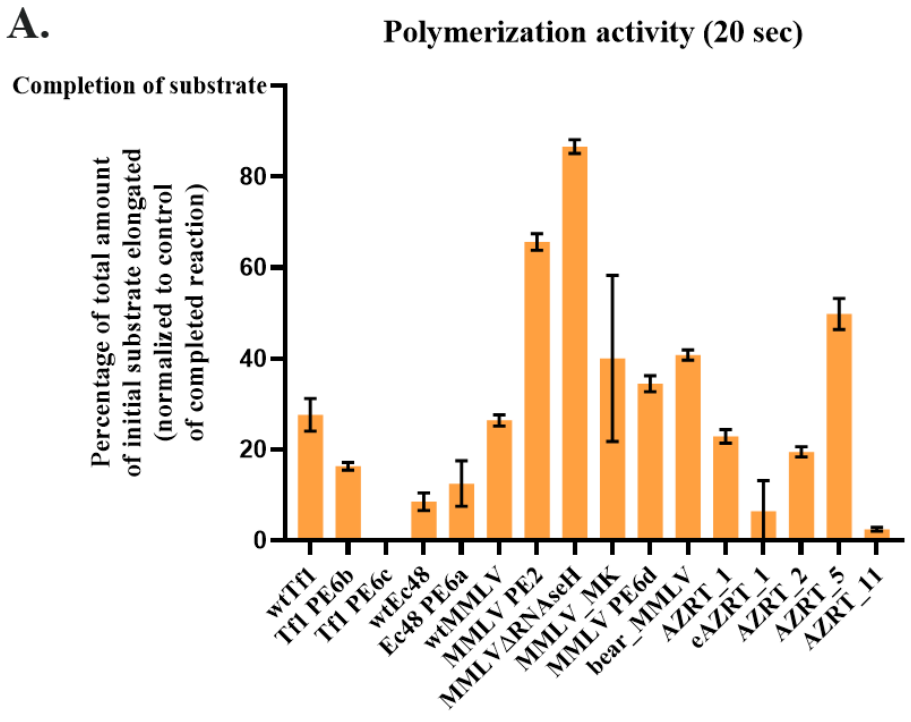
## 4.8 Polymerization Activity

Polymerization activity of the RTs was benchmarked using a polymerization assay with subsequent electrophoresis through a denaturing UREA gel. This measures the velocity of reverse transcription. The assay was first optimized through testing of variables such as substrate concentration, dNTP concentration, reaction time and substrate structure. Finally, it was decided to use the same substrates as for the

binding assay (*figure 4.11A*), however, the FAM fluorophore was exchanged for an ATTO647 fluorophore for higher sensitivity on the gel. The Kds from the binding assay were used for determining a productive protein concentration to reach binding equilibrium. It was determined that 1 $\mu$M of protein should be excessive enough to push equilibrium toward binding of all templates, and for comparability this concentration was used for all RTs. Though, for the AZ RTs this concentration might be too low to push the equilibrium toward complete binding as they seem to have lower binding affinity.

Through time course polymerization used for finding an optimal reaction time to capture the elongation with high resolution, it was determined that 20 second reactions should be enough to distinguish the RTs' performances. Some of the investigated RTs had very high polymerization activity and completed the reaction with longer reaction times by extending all the available templates.

As a negative control, the substrate without protein was used, while ThermoFisher's Multiscribe qPCR kit was used as a representation of a completed reaction, as a positive control. The amount of template extended by Multiscribe in the completed reaction was used to normalize the extended substrate by the other RTs as explained in *section B.4*.



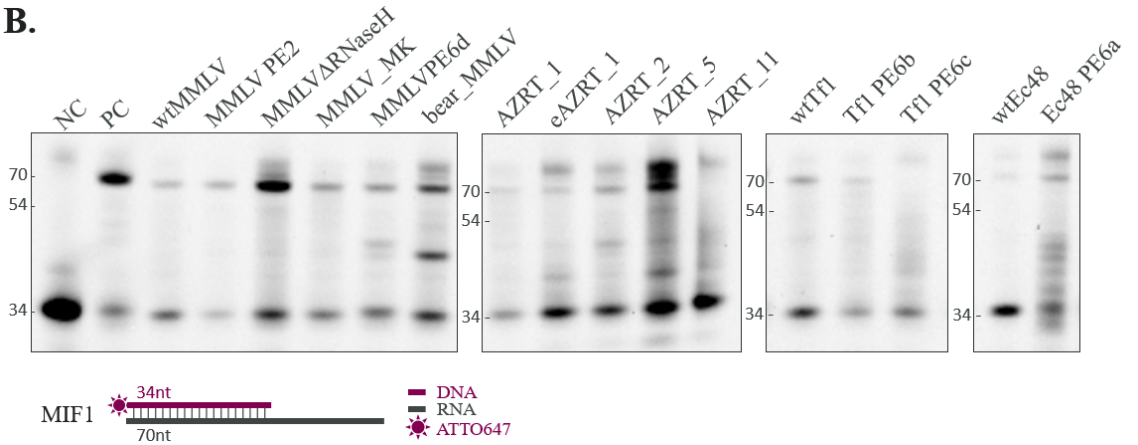



**Figure 4.15: A.** Percentage extended substrate during 20 second reactions. (SD, n=3) **B.** Pooled replicates (n=4) of the polymerization assay.

The result of the polymerization activity assay shows that the RTs have a wide range of velocities, from 0% to approximately 85% extended substrate in 20 seconds (*figure 4.15A*). The standard deviation of the result is relatively low, indicating that most RTs reverse transcribe the substrate in a reliable manner achieving highly reproducible results. However, MMLV_MK and the eAZRT_1 both seem to have more unpredictable activity, creating different products in the replicate samples (*figure B.5*C and D), resulting in higher standard deviations. The polymerization assay shows that the truncated MMLV almost completes the reaction in 20 seconds, extending more than 85% of the available substrate (*figure 4.15A*). Next is MMLV PE2, which extends almost 70% of the available substrate. The wtRTs extend from 10% to about 30% of available substrate, while bear_MMLV extends about 40% of available substrate. The slowest RT is Tf1 PE6c that does not extend any of the substrate in the 20 second reaction. Through the previously mentioned time course assays used for assay optimization, it was confirmed that the Tf1 PE6c needs longer reaction times to reverse transcribe the substrate into the completed product. These results imply that RTs with high editing efficiency can have anything between slow to fast polymerization, and the same is true for the RTs with low editing efficiency.

The polymerization assay also showed what intermediate products that the RTs created. The result of pooled replicates from 20 seconds polymerization reactions is shown in *figure 4.15B*. Some RTs display clear tendencies to pause at certain lengths, for example bear_MMLV and MMLV PE6d, possibly indicating some processivity issues. Some of the AZ RTs also have a lot of intermediate products that might be a result of low processivity, however, might also be a product of the equilibrium not being successfully shifted towards complete binding initially, due to lower binding affinity for the AZ RTs. Ec48 PE6a shows a smear that extends below the original substrate, suggesting substrate degradation through nuclease activity.

Through the polymerization assay, it was observed that the band signal was consistently weaker for some RTs (*figure 4.15B*), especially MMLV PE2 and wtMMLV,

even though initial substrate concentration were identical to other RTs. Some band smears were also observed in reactions with Ec48 PE6a. These observations converged in a hypothesis about some samples degrading the substrates, either the RT by itself, or from contamination by nucleases. This was investigated by incubating the RTs together with the substrate and observing band intensity and potential degradation products (*figure B.13*). This was confirmed to be the case for Ec48 PE6a on both substrates. Furthermore, some nuclease activity was detected in the MMLV family RTs as well, but only on the GAPDH substrate, including MMLVΔRNaseH which are missing the RNaseH domain with intrinsic nuclease activity. These results suggest some nuclease contamination.

## 4.9 Analysis of RT Property Profiles

To explore whether any of the investigated properties were correlated with TE activity, each property was first studied separately. Then radar plots were used to create property profiles that would enable comparisons of the set of properties of selected RTs. First, TE efficiency was investigated in relation to thermostability by plotting the TE in *figure 4.4* against melting point (Tm) in a scatter plot to visualize any pattern or correlation between the two properties (*figure 4.16*).

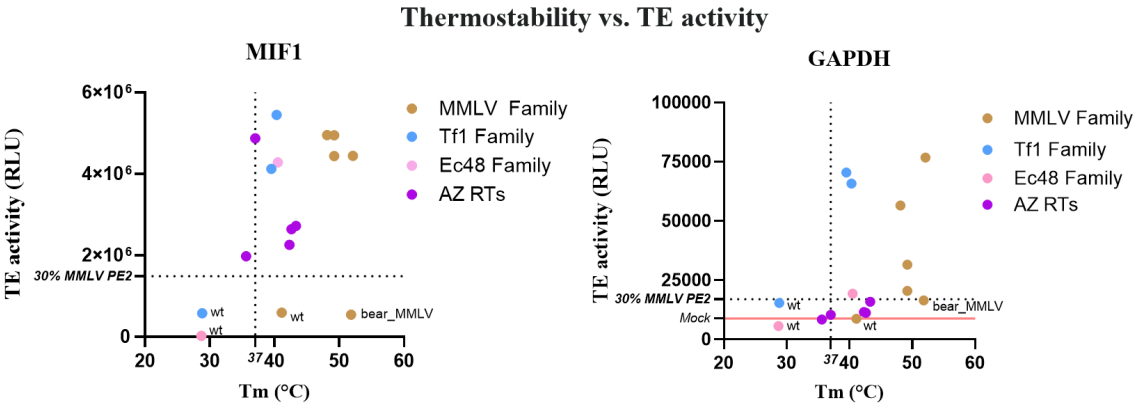



**Figure 4.16:** TE efficiency plotted against melting point (Tm) at each target site.

When plotting TE activity at the two target loci against melting temperature (*figure 4.16*), the scattering of the data points does not indicate a close correlation between stability and TE efficiency. A line has been put at 37°C for easy reference to physiological conditions. Among the editors with low editing efficiency, there is a wide range of melting temperatures from 27 up to 52, which indicates that stability alone does not create an efficient TE. However, there is also no RT that could be considered good at TE with melting temperatures below approximately 35°C which suggests that there might be a threshold to overcome for the RT to even be able to perform TE at all. wtMMLV for example is very interesting in that sense, since it has a considerable level of stability, the same as the engineered PE6b and PE6c RTs, but is unable to efficiently install edits. Suggesting a weakness relating to another

quality. The data suggests that the level of stability does not determine whether the RT will be effective as part of a TE, though, improving thermostability in wtRTs does seem to correlate with them improving in TE applications.

Interestingly, this data suggests a potential improvement of the Tf1 family. Tf1 RTs already performs as well as the MMLV family for many edits, however, are significantly less stable. An improvement in stability might lead to an even higher increase in TE efficiency.

Next, TE efficiency was plotted against binding affinity (Kd) on both target sites (*figure 4.17A*).

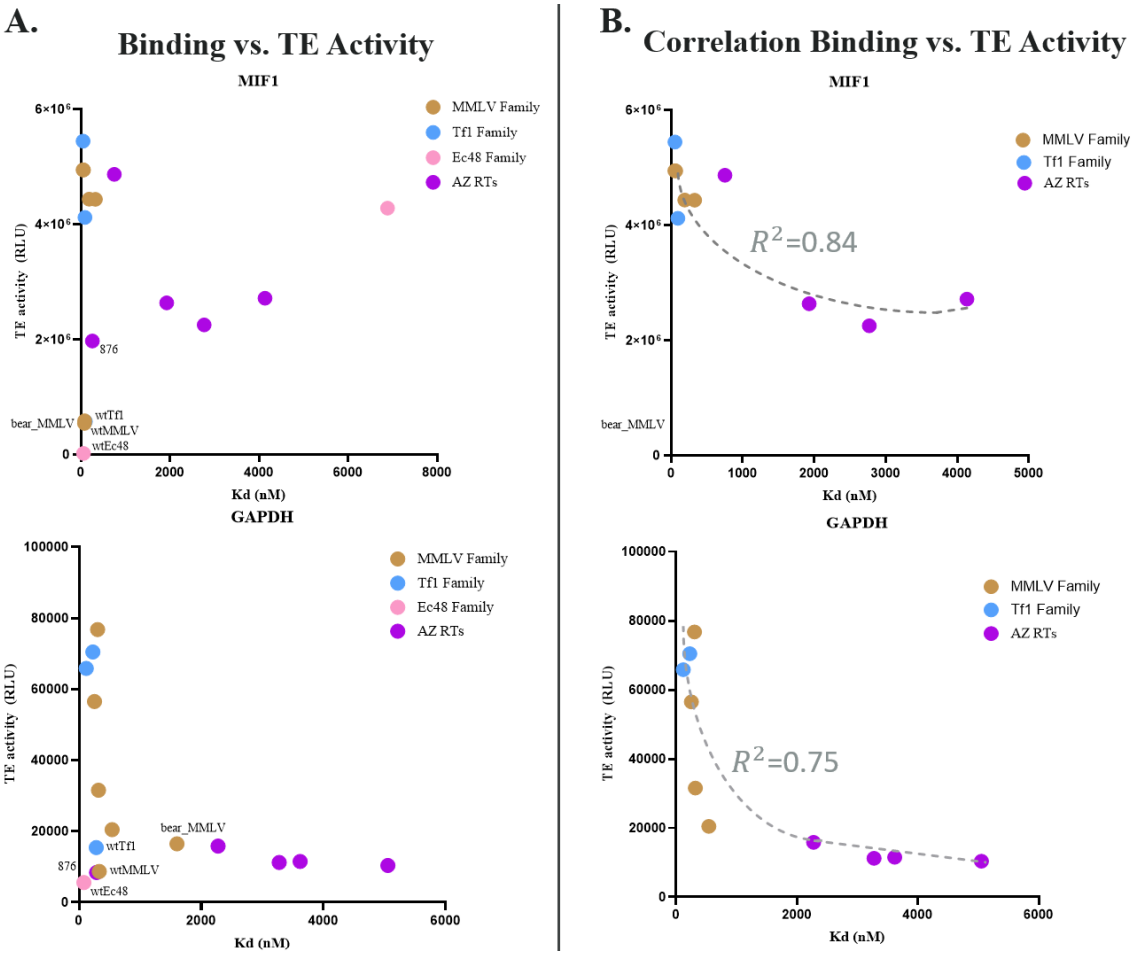



**Figure 4.17:** **A.** TE efficiency plotted against Kd at each target site. (CMV promoter) **B.** A non-linear regression is fitted to the same data set as in A., excluding wtRTs, PE6a, AZRT_5 and bear_MMLV as outliers.

When Kds are plotted against TE activity using the CMV promoter to express the RT (*figure 4.19A*), one particular group of RT sticks out. These are the RTs that have very low TE efficiency: wtRTs, bear_MMLV and AZRT_5. These all group together close to origin due to their low Kds and low editing activity. Removing

these RTs from the plot together with Ec48 PE6a suspected to be an outlier, an inverse non-linear relationship between activity and Kd appears, where increased Kd reduces TE activity (*figure 4.17B*). The removal of inefficient RTs for TE before the analysis, suggests that the inverse relationship is true only if there are also some other basic qualities present in the RT to allow them to work in a TE setting. All of the wtRTs and bear_MMLV lack these basic qualities and are therefore not able to gene edit efficiently even though they interact strongly with the DNA/RNA substrates. AZRT_5 might possess similar qualities as the wtRTs since it groups with them. The regressions suggest that it is advantageous for TE efficiency if the RT binds strongly to the substrate. However, as with stability, these results disprove a direct correlation between binding and TE activity.

Lastly, TE efficiency was plotted against polymerization activity (%) on both target sites (*figure 4.18*).

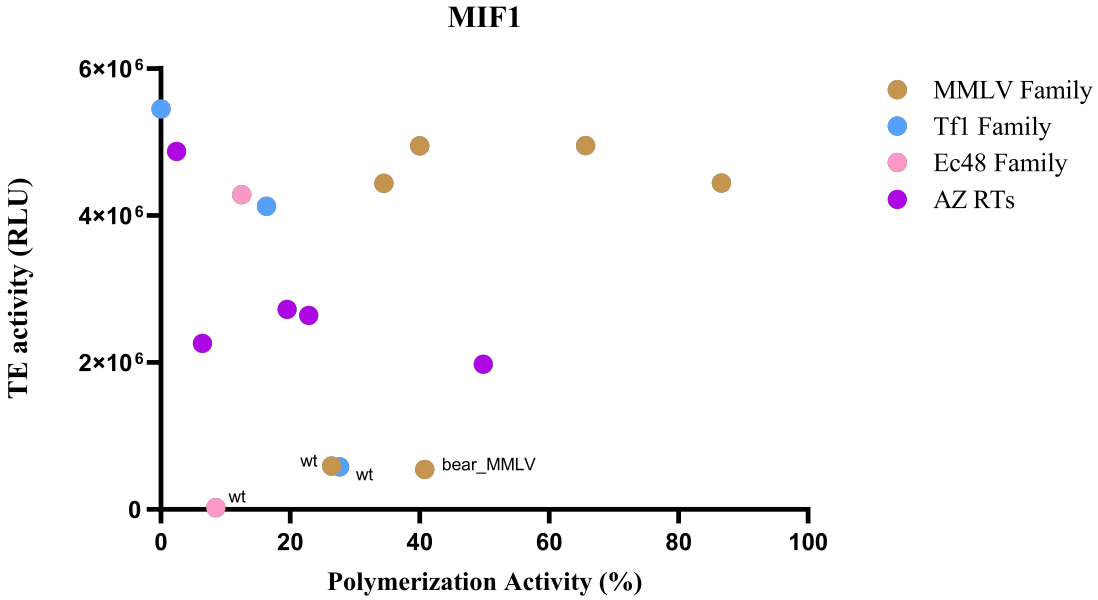



**Figure 4.18:** Average editing efficiency at MIF1 locus plotted against polymerization activity of MIF1 substrate.

When plotting TE efficiency at the MIF1 locus against polymerization activity of the RTs using the MIF1 substrate (*figure 4.18*), there is no correlation between the two. The data suggests that TE efficiency is completely independent from the velocity of polymerization of the template. This does make sense in the context of TE, since the edits incorporated during TE are usually of a size smaller than 50 bp, and at this size, velocity might not be the most important factor for successful editing. More relevant are probably properties relating to how well they perform the elongation, meaning how well they bind, process, and handle different substrates and secondary structures of the RTT rather than the speed that the can install this short stretch of DNA.

These correlation analyses indicates that optimizing RTs for TE might be a multi-factorial problem. It seems thermostability is only important if other qualities are present; Substrate binding strength only matter if the RT is proficient at editing; And many low efficiency RTs have potential issues with solubility. To investigate the relationship between these properties and editing outcomes, radar plots were created to compare the properties of certain groups of RTs: efficient and inefficient RTs for TE, and the RT optimized for qPCR. To determine which RTs to include in which group a 2D plot was created showing TE efficiency at each target site on separate axes (*figure 4.19*).

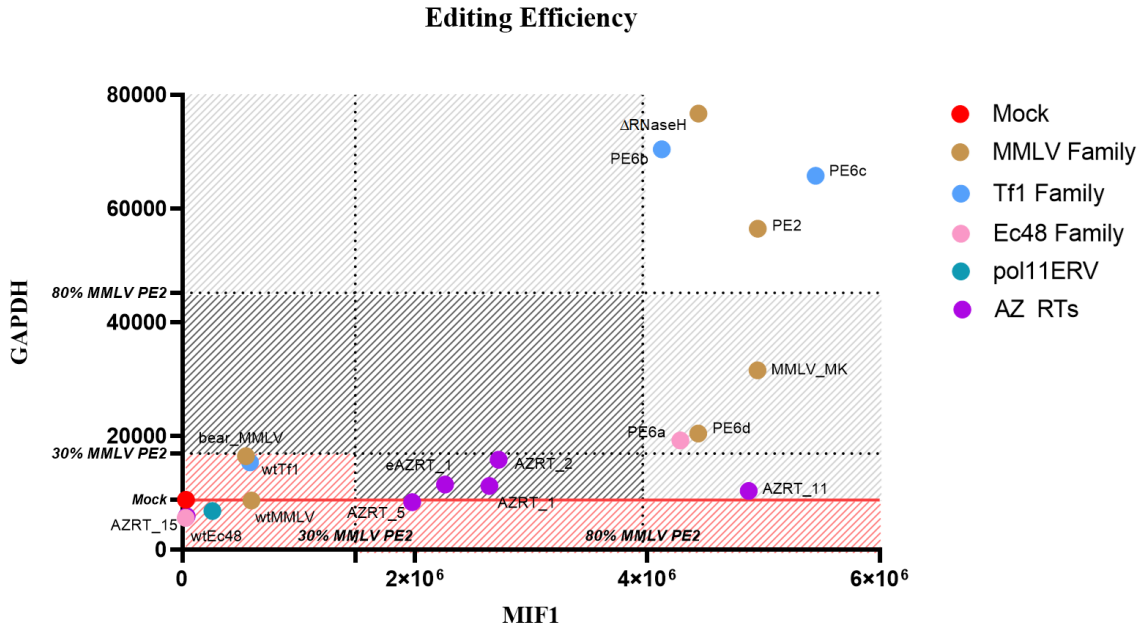



**Figure 4.19:** Average editing efficiency at GAPDH locus plotted against average editing efficiency at the MIF1 locus. Cut-off lines are defined as 80% of the TE activity (RLU) of PE2 at that target site. Numbers on the axes are RLU normalized to GFP/Phase Area. The red area indicates areas where RTs are considered as having no TE activity. Dots are average of 1 to 4 replicates where the RT have been expressed with CMV promoter.

To determine the reference points for which RTs would become the baselines for each group, grid lines were drawn on the plot to define high efficiency and low efficiency (*figure 4.19*). An RT would be considered highly efficient at editing on the target site if it had at least 80% of the MMLV PE2's TE efficiency at that site. And an RT with less than 30% of the MMLV PE2's editing efficiency would be considered to have low efficiency on the target site. This resulted in nine boxes that have been marked with dotted lines in *figure 4.19*. The bottom left corner highlights RTs that are highly inefficient ("bad" RTs). The top right corner surrounds RTs that are very efficient at both target sites ("good" RTs), and the middle boxes fits RTs that are intermediate TEs or very site specific. Many of the AZRTs fell into the intermediate and site specific category, while PE6b, PE6c, MMLVΔ and PE2 were

defined as highly efficient for TE. For contrast, I selected the wtRT as references for RTs that have low TE efficiencies, and bear_MMLV was included as the RT has been engineered, and works well for qPCR which it was engineered for, but does not possess properties that are favorable for TE efficiency.

For most properties I used the value from the assay directly. However, for site dependency a scoring system was applied based on the boxes in *figure 4.19*. The low efficiency RTs in the lower left box were defined as not site specific (as they have low efficiency on both target sites), top right were also defined as non-specific (as they have high efficiency on both target sites). Both of these boxes received a 0 as score. The top left and bottom right were instead defined as site dependent and received the score 1 for the analysis.

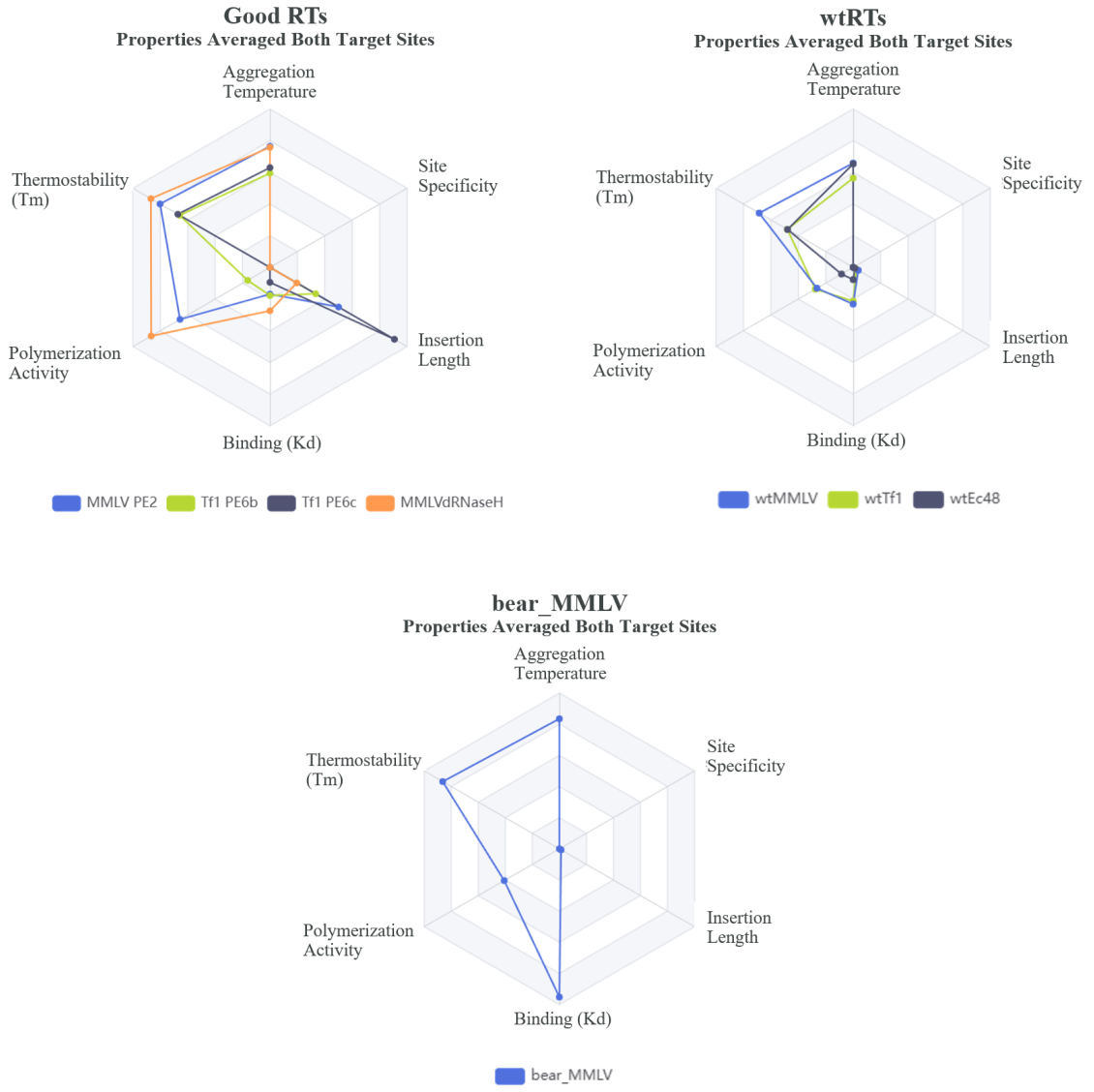



**Figure 4.20:** Individual trait profiles of good RTs, wtRTs and bear_MMLV. High Kd is considered negative.

*Figure 4.20* shows the property profiles of the highly efficient (good) RTs, wtRTs and bear_MMLV RT respectively. The figure shows general behaviors of the RTs by averaging properties at the two target sites. The graphs show trends among the different groups, the good RTs possess properties that are favorable for TE efficiency, while the wtRTs illustrate what the starting point was. The comparison between the two groups shows what properties were improved and to what extent to make the RTs work in a TE context. The bear_MMLV RT on the other hand, shows other unfavorable traits for TE efficiency. This is another direction of engineering wtMMLV that does not work for TE.

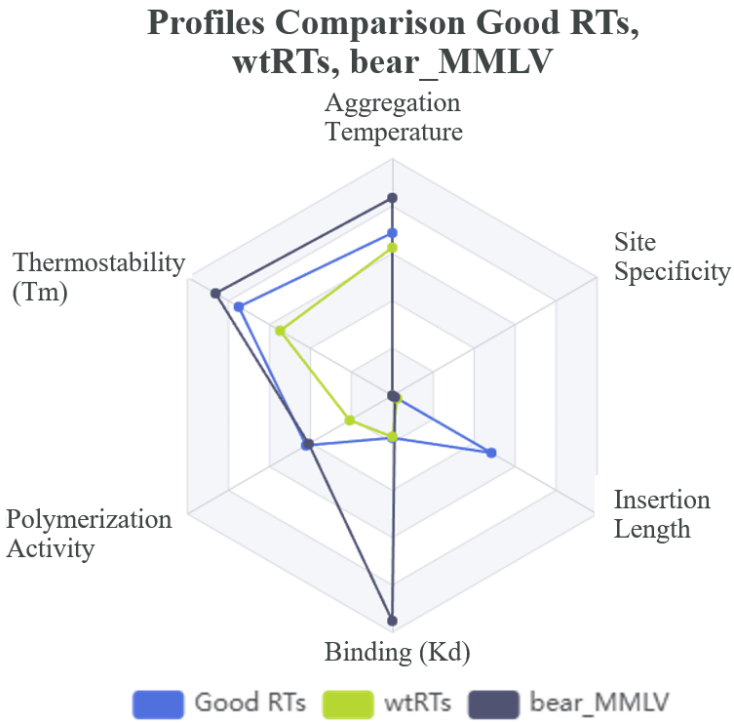



**Figure 4.21:** Property profiles of good RTs, wtRTs and bear_MMLV. "Good RTs" show the average value of each trait from the RTs MMLV PE2, Tf1 PE6b, Tf1 PE6c and MMLVδRNaseH. "wtRTs" show average value of each trait from the RTs wtMMLV, wtTf1 and wtEc48. Average of both target sites. High Kd is considered negative.

When averaging the RTs in each group, the trends become clearer (*figure 4.21*). In general it seems the efficient RTs have relatively high thermostability, aggregation temperature, strong binding interaction and insertion length. The efficient substrate interaction, both through binding and insertion length, suggests high substrate affinity. wtRTs are in general less stable and more prone to aggregation. Though they also have strong binding to DNA/RNA substrates and thus seem to interact with high affinity to the substrate, however, they are not good at reverse transcribing longer substrates. Bear_MMLV have a strong combination of stability and aggregation resistance, but shows low affinity to at least one of the substrates. The main differences between these groups are stability, aggregation, and substrate affinity.

Comparing good RTs and wtRTs, both have similar Kds. What separates them is the stability and solubility of the protein. My hypothesis is that for substrate affinity to matter in TE contexts, the RT needs to be stable enough not to aggregate or degrade in the cellular environment. Thus, improving solubility and thermostability should be the first priority for RT optimization. However, the traits possessed by bear_MMLV suggests that there might be an optimal interval where too stable and resistant to aggregation, instead prevents the enzyme from preforming successful editing. The next step would be to improve substrate affinity.

## 4.10 Engineering

To optimize AZ RTs and confirm acquired hypotheses regarding the effect of properties on TE efficiency, the rational engineering of the RTs aimed at improving solubility, thermostability and substrate affinity. Solubility arose as a potential issue during the purification, as RTs known to have low TE efficiency had trouble with solubility. Thermostability was determined to have the potential to boost editing based on the gathered data and substrate affinity and interaction seem to be critical for successful installation of RTTs.

The approach was to use a combination of rational engineering and AI tools. The actual procedure is explained in detail in *section 3.8*.

### 4.10.1 Gen I

Two of the AZ RTs of interest for engineering were eAZRT_1 and AZRT_11. Both of these enzymes showed potential as editors with some improvements, and the biochemical and biophysical studies of these two proteins indicated areas which could be improved through engineering, see *figure 4.22*.

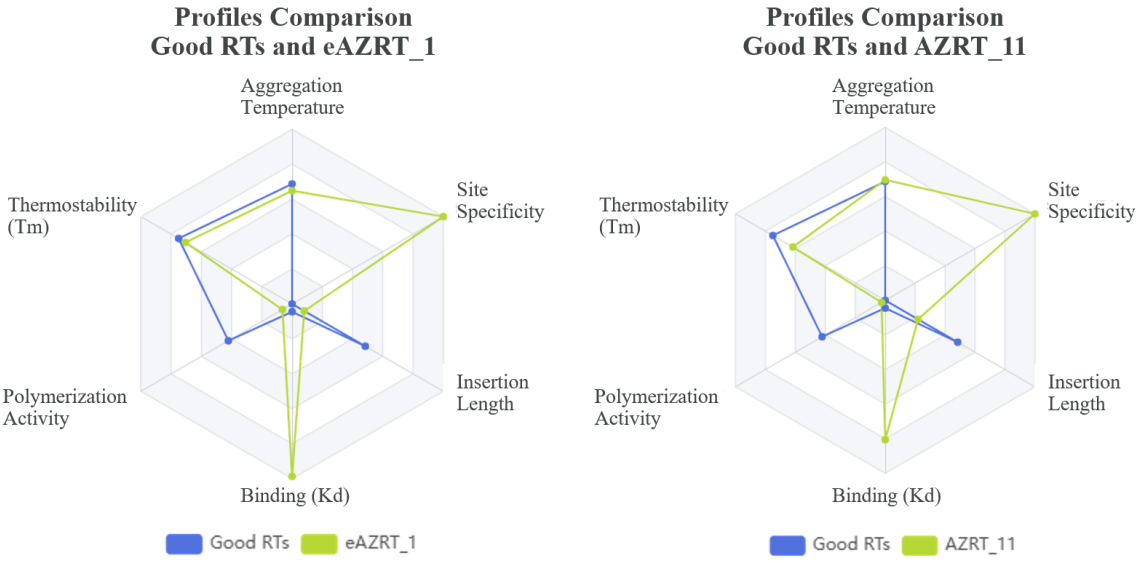
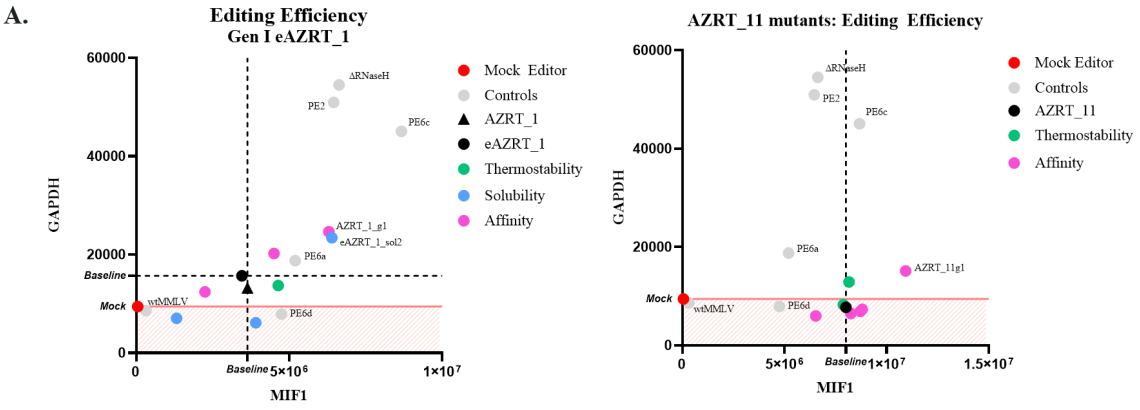



**Figure 4.22:** Property profiles of good RTs and eAZRT_1 and AZRT_11. High Kd is considered negative.

The main issue affecting both eAZRT_1 and AZRT_11 was low affinity to the substrates indicated by high Kds (*figure 4.22*). The RTs also had problems with longer substrates, possibly also processivity issues indicated by prematurely terminated products being created during polymerization. eAZRT_1 had less problems with thermostability and aggregation, though, it was hypothesized that improving these properties could boost the editing efficiency further. The same rational was applied to AZRT_11 as well, with emphasis on thermostability as the RT seems to fall short in this category in comparison with efficient RTs. It was also identified during the purification, that both of these RTs might suffer from misfolding due to some potential solubility issues, however, both RTs show adequate resistance against aggregation in the turbidity assay.

All mutants were screened for TE activity using the saturated 2-component system at both the GAPDH and the MIF1 loci. After an initial screen, the most promising candidates were selected and benchmarked on the two loci, to obtain more robust data. In *figure 4.23A* the results of the editing on the two loci have been plotted against each other, and the rational behind the mutant highlighted in either category: affinity, solubility, and thermostability. Some of the affinity point mutations that were transplanted improved editing while some others negatively affected the enzymatic activity. For eAZRT_1 the best variants that improved efficiency were a cluster of mutations that improved affinity and structure (AZRT_1_g1) and one of the soluble proteinMPNN variants (AZRT_1_sol2). For the AZRT_11, the most effective engineered variant possessed a single mutation affecting DNA-binding, and enhanced editing across both target sites, AZRT_11g1. Another mutant with a transplantation affecting DNA affinity also improved editing, but only on the MIF1 locus.



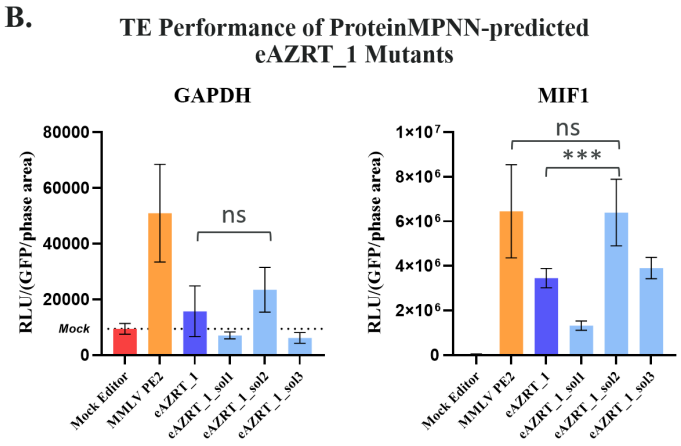



**Figure 4.23: A.** TE efficiency of GEN I mutants. 2-component system overexpression the RT with CMV promoter. The dots are color coded to visualize what the rational behind the introduced mutation was. **B.** TE efficiency of ProteinMPNN-predicted mutants of eAZRT_1. Significance tested with student's T-test assuming the same variance for the groups. (SD, n=6 to 11)

The results in *figure 4.23A* indicates that the single mutations from ThermoMPNN to improve thermostability did not result in any activity improvements. Possible reasons might be that the pipeline does not take enzyme function and domains into consideration, thus it is possible that conformational changes or flexibility were restricted by the introduced mutations. Though, one structural mutation for each enzyme improved editing somewhat at one target site.

For the AZRT_11 there were no mutants engineered from the ProteinMPNN predictions that maintained the enzymatic activity, however, for eAZRT_1 there were two candidates from Soluble ProteinMPNN that maintained TE activity (*figure 4.23*B). Strikingly, the eAZRT_1_sol2, improved TE efficiency at both target loci, enhancing editing efficiency to levels comparable to MMLV PE2 at the MIF1 locus. The effect was not significant at the GAPDH site. These results further suggest that the eAZRT_1 had problems with solubility and stability, which reduced its editing efficiency *in vivo*.

Two other enzymes to be engineered were AZRT_15 and Ec48 PE6a. These were selected for engineering as it was hypothesized that engineering their backbone to improve solubility and structure, could improve the efficiency of these enzymes. Both of the RTs lacked in terms of stability as suggested by the difficulty to purify the RTs. Both misfolded during expression in *E. coli*, and formed inclusion bodies. Some rational engineering to improve the structure was also made to AZRT_15, by for example exchanging some surface cysteine that might increase the risk of aggregation and misfolding. The mutants were screened on both MIF1 and GAPDH loci (*figure 4.24*). For the AZRT_15, the rational engineering was not very effective. The soluble ProteinMPNN variants had a more positive impact on TE efficiency than the rational engineering, however, the successful soluble ProteinMPNN vari-

ants did still not improve editing significantly. Improving structure is probably not enough to rescue this RT. There are probably other properties such as weak binding interactions, that prevents this RT from performing TE even when stability and solubility is improved. This further supports the hypothesis that stability, solubility and binding have to be optimized together to improve TE efficiency. Though, without biochemical and biophysical data about the AZRT_15 this cannot be confirmed in the case of this RT.

In the case of Ec48 PE6a, improving the backbone of the enzyme with ProteinMPNN predictions significantly improved the TE efficiency on the MIF1 target locus. On the GAPDH site however, there is no improvement.

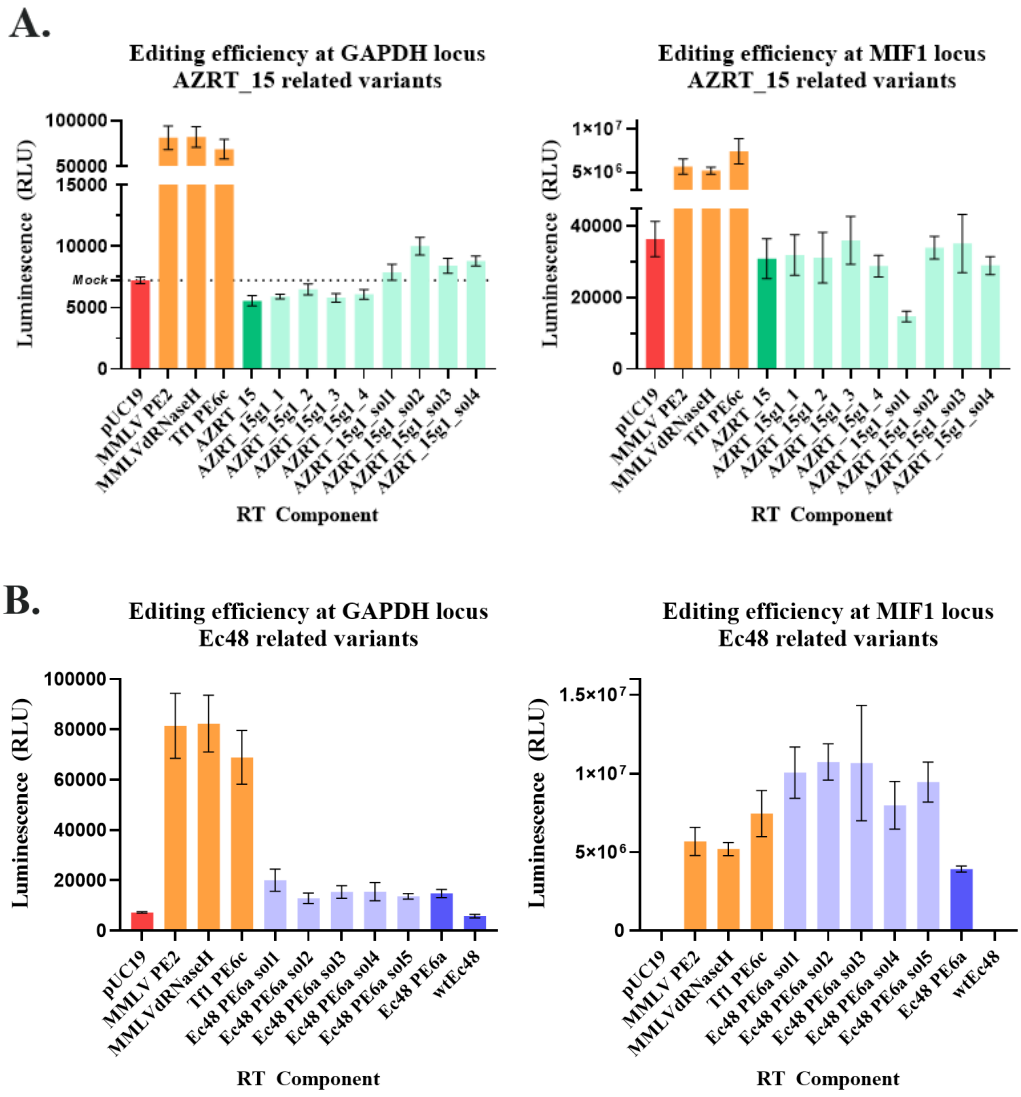



**Figure 4.24:** **A.** TE efficiency of AZRT_15 GEN I mutants. **B.** TE efficiency of Ec48 PE6a ProteinMPNN solubility mutants.

### 4.10.2 Gen II

When engineering the second generation of eAZRT_1 and AZRT_11, the best candidates from the first generation were picked to provide the base to stack mutations onto. For AZRT_11, the best variant from Gen I (AZRT_11g1) was used as base for stacking additional mutations. The second best affinity mutation was combined with AZRT_11g1, and to improve the thermostability further, the best thermostability mutant was combined with three other mutations suggested by ThermoMPNN to be favorable in combination with the mutation in question, to create three new double mutants. Additionally, rational engineering was done to introduce structural mutations possibly affecting either stability, flexibility or solubility. Rational engineering was also done to improve affinity by modifying the charge and polarity of DNA-interacting residues. Additional mutations were introduced to the AZRT_11g1 to affect the structure and DNA binding. In the second generation, related proteins were studied to identify the most conserved residues that AZRT_11 did not possess. These were also introduced to AZRT_11g1, most being structural changes. The engineered AZRT_11 variants were screened at the two target loci together with the best variants from the first generation and the wtAZRT_11 (*figure 4.25*).

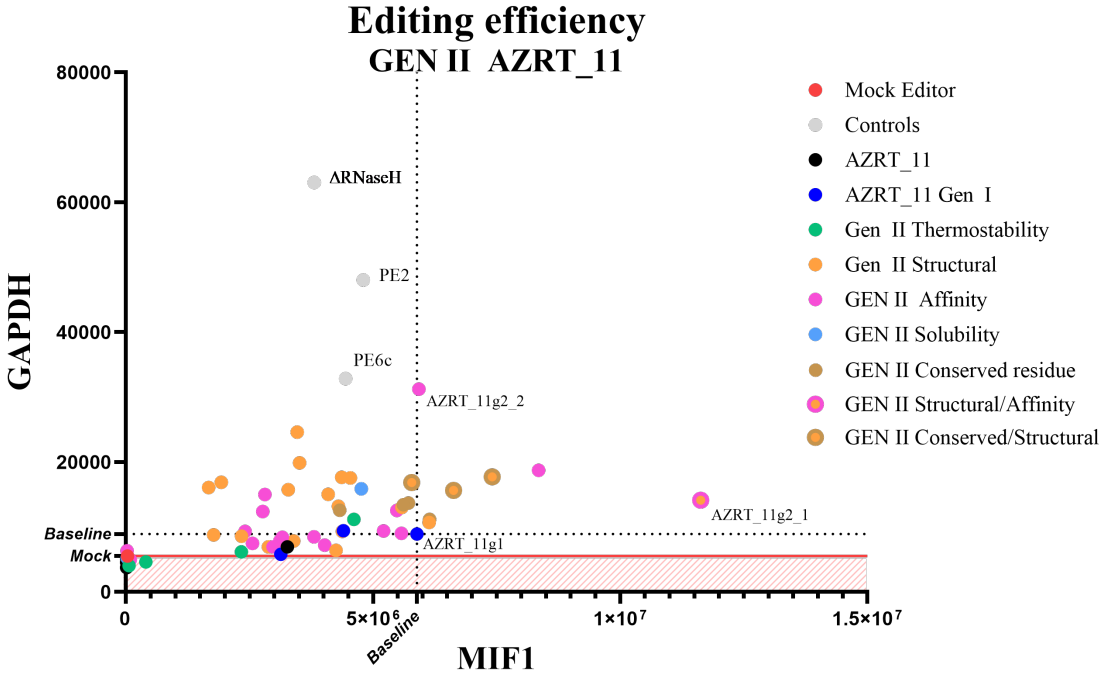



**Figure 4.25:** Engineered AZRT_11 second generation screening at the GAPDH and MIF1 loci. The dots are color coded to visualize what the rational behind the introduced mutation was.

The results show that most mutants only improved editing efficiency marginally at one or both target sites (*figure 4.25*). Targeting thermostability using ThermoMPNN seems to be relatively futile, however, there is one double mutant that improved editing compared to the wtAZRT_11, but still performed worse than the

AZRT_11g1 from the first generation. In general, it seems improving structural stability and flexibility increases editing efficiency. Targeting affinity also proved to either yield no effect or improve editing efficiency immensely. Introducing the conserved residues from related proteins, improved editing in most cases. In general, the best candidates from the second generation are all combinations of AZRT_11g1 and a mutation targeting affinity. One of them interacts directly with the bases in the DNA/RNA complex close to the active site, while another improves flexibility of the thumb domain while exchanging a DNA-interacting residue. This variant, AZRT_11g2_1, seems to have very high sequence specificity as it is about 2.4 times as effective as MMLV PE2 on the MIF1 locus, but very low activity on the GAPDH locus in comparison with MMLV PE2. Interestingly, the best variant on the GAPDH site, AZRT_11g2_2 did not improve on the MIF1 locus, however, the improved DNA interaction significantly enhanced editing efficiency at the GAPDH locus.

In contrast to other RTs, the AZRT_11 does not seem to tolerate the Soluble ProteinMPNN predicted changes to its backbone. This reduces its TE efficiency to zero (the dots are hidden behind other mutants by the mock editor in *figure 4.25*). Even though all important domains and DNA-binding groove were liberally locked when submitting the structure for ProteinMPNN predictions, and the final variants passed through several manual quality controls and scored high in Boltz-2 predictions with the substrate, it seems the altered shell severely affects the TE activity of this RT. This suggests some important site or structural element, previously missed present on the surface of AZRT_11 that is essential for the TE activity.

In the case of eAZRT_1, the ProteinMPNN-predicted eAZRT_1_sol2 was used as a base to combine other mutations from previous generations. A few combinations were designed and screened for TE efficiency at the GAPDH and MIF1 loci (*figure 4.26*). The results showed that the second generation either improved, or did not affect editing efficiency significantly compared to eAZRT_1_sol2. The best mutant performed comparable to the MMLV PE2 on both target sites, and was a combination between eAZRT_1_sol2 and AZRT_1_g1.

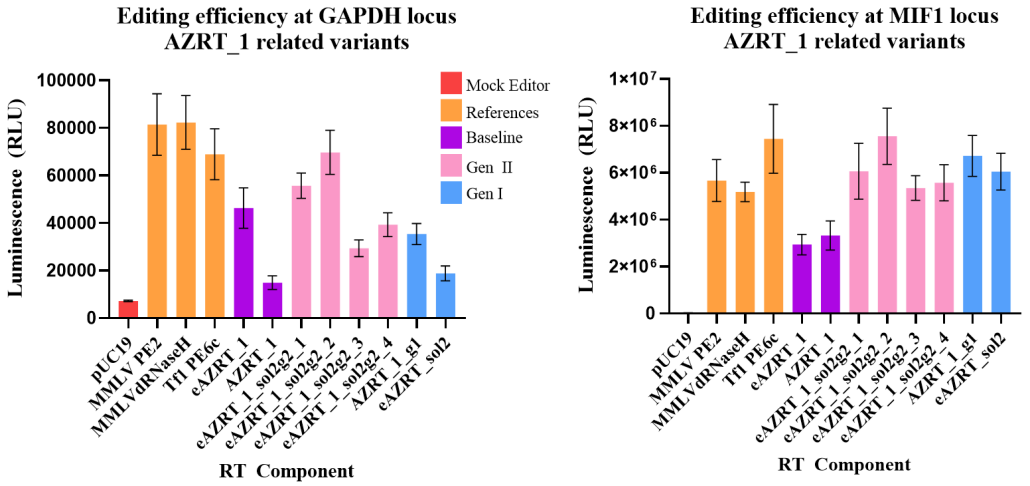



**Figure 4.26:** Engineered AZRT_1 second generation screening at the GAPDH and MIF1 loci.

## 4.11 Limitations and Future Direction

The results of this project suggests that the three RT properties thermostability, solubility and DNA/RNA affinity have a synergetic effect on TE outcomes in cells which was confirmed through complimentary cellular and mechanistic studies and engineering cycles. However, one limiting aspect of this study is that it has not been confirmed that the mutations targeting certain properties actually had the intended effect. Future studies should prioritize mechanistic validation by investigating the biochemical and biophysical properties of the best engineered RTs to confirm the reasoning in this work. Furthermore, these engineered RTs' template length capacity should also be investigated to validate how these mutations influence insertion length which might reveal engineering approaches for enabling longer edits.

Future biochemical and biophysical studies should also consider solubility in mammalian cells, and processivity. Due to time constraints, solubility could not be included in the selection of properties studied in this work after it was discovered that solubility might restrict the activity of multiple RTs. So, even though solubility was improved through engineering with positive effect, solubility levels of the RTs remains to be properly investigated, especially since the AZRT_11 could not be improved through the limited number of tested variants with remodeled surfaces. Solubility must be determined to explain the case of AZRT_11: Is there an optimal range of solubility as was suggested for thermostability, or are there some RTs that have important elements on the surface? For processivity, this property should be thoroughly investigated as previous work have established this property as a significant player in template installing said to influence scaffold incorporation, length that the RT can install and wether the RT can handle secondary structures of the RTT. To my knowledge, no proper processivity investigation of TE RTs has been made so far, meaning these theories are mostly speculative and illustrating the need for proper investigation to see if these observations are related to processivity in

reality or could be connected to other properties.

Future studies should also consider how representative the results are and how well they will translate to therapeutic applications. More target sites and cell types have to be investigated. Furthermore, the dose has to be taken into consideration. Usually, the effective doses *in vivo* are much lower than what is used *in vitro* due to delivery method and transfection efficiency of the cells among other factors. The dose-response assay of the RTs showed what the performance of the RTs might be with a dose more similar to *in vivo* conditions, showing that some RTs perform better under these conditions than with the high doses. The reason for the improved performance is unclear, but might be related to stability. For future screening purposes it might therefore be beneficial to use weaker promoters to identify the RTs that perform well under these conditions. Using weaker promoters could also allow for better resolution in the HiBiT assay which could improve the screening process and enable identification of significant performance improvements. It is possible that the current setup of the assay with overexpression of the components overestimates some RTs performances, and underestimates others, resulting in misleading results that reduce the efficiency of the engineering as focus is put on the wrong candidates. Using the HiBiT reporter assay allowed this project to proceed fast with short turn-around-times which worked well for screening and *in vitro* testing of engineered proteins. However, using a reporter assay also meant a loss of some information, including specific editing and indel ratio. Moving forward towards *in vivo* applications, sequencing data of the edits through NGS needs to be collected to determine the RTs' fidelity when editing.

Next, since site specificity arose as an intrinsic property of some RTs, it would be illuminating to investigate what factors influence this and how to engineer specificity. Aided by mechanistic studies of the engineered RTs that harbors single mutations that only had a substantial effect on one target site, these studies could reveal important design considerations for therapeutic applications where specificity is essential. As many RTs in this thesis were revealed to favor one target site over another, it might be beneficial for therapeutical applications to engineer and optimize RTs for specific targets. The results from the engineering efforts of the AZRT_11 seem to indicate that highly specialized RTs could achieve higher editing efficiencies than universal RTs.

The last step would be to create the next generation of engineered RTs by combining promising mutations. Combinations of the best candidates to create competitive TEs, as well as combining mutations that had a significant effect on one target site to see if these can be combined to create an RT that is less site specific.

# 5

# Summary and Conclusion

This project's systematic characterization and engineering of reverse transcriptases revealed critical principles for developing RTs suitable for therapeutic templated editing applications. Through previous work, it has been shown that RT's intrinsic properties likely set the ceiling for how robustly RTTs are copied and installed during TE. Building on this foundation, this thesis project identified the three properties thermostability, solubility, and substrate affinity as determining factors for TE efficiency, and demonstrated that these properties must be collectively optimized to achieve meaningful improvements in cellular editing efficiency.

First, thermostability emerged as the foundational requirement for RT function in templated editing applications. The data demonstrated a threshold effect around 35°C, below which RTs failed to achieve meaningful editing efficiency regardless of their other properties. In contrast, all engineered variants that demonstrated good editing performance exceeded this threshold and had higher thermostability than the wt counterparts. Improvement in thermostability probably directly correlates with the ability to maintain catalytic activity over the extended period of time required for efficient templated editing while exposed to the challenges of the cellular microenvironment. RTs that lack sufficient thermostability presumably unfold or aggregate under these conditions, losing activity before they can complete the reverse transcription of the edit. However, thermostability alone proved insufficient for successful editing, as demonstrated by the bear_MMLV variant, which exhibited high stability yet remained ineffective for TE. This observation suggests that excessive stabilization might introduce rigidity that impairs the conformational flexibility required for substrate engagement in the cell, suggesting an optimal stability range for RTs. ThermoMPNN was used to predict favorable stability mutations, however, proved to be quite ineffective at suggesting productive mutations for this specific application. In contrast, other rationally engineered, or transplanted residues aimed at improving stability and structural integrity proved to be very effective at improving TE outcomes in cells. In summary, thermostability is necessary but not sufficient for TE efficiency. Optimizing this property involves keeping stability within an optimal range so as not to make the protein to rigid.

Secondly, RT solubility and resistance to aggregation represented the second critical factor for TE efficiency, and the application of ProteinMPNN for computational

redesign of RT backbones proved remarkably effective at addressing this challenge. Several RTs in this study formed inclusion bodies during expression in *E.coli*, indicating fundamental misfolding and solubility issues that would likely manifest in mammalian cells as well. The AstraZeneca proprietary RTs eAZRT_1, AZRT_1, and AZRT_2 were extracted at extremely high concentrations and formed visible aggregates during purification, with solutions becoming cloudy and viscous. These observations suggested that even when these RTs could be expressed, their tendency to aggregate might limit their effective concentration in cells and potentially sequester them in non-functional complexes. To address these challenges, soluble ProteinMPNN was used to redesign the protein surface to improve hydrophilicity and reduce aggregation-prone patches while preserving the core structural elements and active site architecture. For eAZRT_1, the ProteinMPNN-predicted variant eAZRT_1_sol2 achieved editing efficiency comparable to MMLV PE2 at the MIF1 locus, representing almost a 2-fold improvement over the wild type RT. This computationally designed backbone transformed a poorly performing RT into a competitive editor, validating the hypothesis that solubility was the primary limiting factor for this enzyme. The approach showed similar promise with other RTs, the Ec48 PE6a backbone improvements significantly enhanced editing efficiency at the MIF1 locus, however, the improvements were less pronounced at the GAPDH site, suggesting that solubility improvements alone could not overcome all limitations of this RT family. The ProteinMPNN approach proved particularly valuable because it addressed solubility issues without requiring detailed mechanistic understanding of why specific RTs aggregated.

Despite these successes, the soluble ProteinMPNN approach also revealed important limitations that must be considered in future engineering efforts. The complete loss of activity in AZRT_11 variants containing ProteinMPNN-predicted backbone modifications, despite careful masking of all known functional domains and rigorous quality control, demonstrated that computational predictions can inadvertently disrupt critical but unrecognized structural elements. The specific functions of these features remain speculative. However, since they are present on the surface of the protein, they could potentially have interactive properties to other proteins or molecules, be important for specific conformational dynamics, or involved in proper cellular localization or stability - features that are not apparent from static structures. The contrasting success with eAZRT_1 and Ec48 PE6a compared to AZRT_11, suggests that some RT scaffolds are more tolerant of surface remodeling than others, possibly reflecting differences in their evolutionary optimization or structural architecture, though the limited number of tested variants might also be the reason for the drastic differences. However, the notable enhancements in 3 out of 4 RTs achieved by remodeling the surface residues demonstrates the significance of protein solubility in TE outcomes.

Lastly, substrate affinity of the RT represented the third critical determinant of TE efficiency, and the engineering of this property proved both highly effective and mechanistically complex. The fluorescence polarization assays revealed that wild type RTs generally exhibited low Kd values indicating strong binding to DNA/RNA

substrates, yet these same enzymes showed poor editing efficiency in cells. This apparent paradox was resolved by recognizing that substrate affinity only translates into functional editing when coupled with adequate thermostability and solubility. This finding reinforces the theory of interdependence of the three properties. Among the engineered variants that possessed sufficient stability and solubility, an inverse non-linear relationship emerged between Kd and editing efficiency, where RTs with stronger substrate binding (lower Kd) generally achieved better editing outcomes. The AstraZeneca RTs, which showed higher Kd values, consistently underperformed in cellular editing assays, suggesting that weak binding reduces the probability of productive RT-substrate engagement during TE in cells. Rational engineering to improve substrate affinity through targeted point mutations proved highly effective, particularly when guided by comparison with high-performing RTs. A striking example are the second-generation variants that combined multiple affinity-enhancing mutations with the initial improvements to achieve even more dramatic results, with AZRT_11g2_1 reaching 2.4-fold higher editing efficiency than MMLV PE2 at the MIF1 locus. The mutations that proved most effective targeted residues directly involved in DNA/RNA interaction near the active site or enhanced the flexibility of the thumb domain to improve substrate accommodation. Interestingly, mutations affecting substrate affinity showed strong site-specificity in their effects, with some variants dramatically improving editing at one locus while showing minimal benefit or even reduced activity at another. This context-dependence suggests that substrate affinity engineering can inadvertently introduce sequence preferences or structural constraints that favor certain RTT configurations over others, an effect that must be considered when designing RTs for therapeutic applications where editing a specific disease-relevant locus is the primary goal.

The findings from individual property optimization highlighted the multifactorial nature of RT optimization, which became evident through the iterative engineering process. First-generation improvements that addressed only one property generally produced modest or inconsistent effects on editing efficiency. However, second-generation variants that combined solubility improvements from ProteinMPNN with affinity-enhancing point mutations achieved improvements that substantially exceeded the sum of individual effects. This synergy demonstrates that the three properties thermostability, solubility, and substrate affinity, are interdependent and must be collectively optimized. The interdependence is evident in several scenarios: Attempting to improve substrate affinity in an RT that aggregates in cells proves futile because the enzyme never reaches its target in functional form. Similarly, stabilizing an RT that cannot bind its substrate effectively merely creates a stable but non-functional protein.

By systematically defining the biochemical and biophysical properties of RTs that determine TE efficiency, this project directly addressed the knowledge gap between the mechanism of TE and cellular editing outcomes. These findings establish clear design principles for engineering RTs with the goal of improving TE efficiency. However, to determine how the results translates to therapeutic applications where the effective concentration and genetic targets are different, further studies are needed

taking different cell types and targets into consideration. As the assays in this project showed that the specific target had a significant effect on editing efficiency for some RTs, it would also be necessary to investigate the sequence specificity and whether this information could be leveraged for engineering purposes when developing therapies for specific targets. Finally, the design principles principles defined in this thesis could reduce costs and time by enabling informed rational engineering that targets properties that actually make a difference for editing efficiency. This approach represents a shift from empirical and speculative engineering toward rational design, thus accelerating therapeutic development.

# Bibliography


1. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, Murphy D, Cam YL, and Rath A. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet 2020; 28. DOI: 10.1038/s41431-019-0508-0
2. Roth TL and Marson A. Genetic Disease and Therapy. Annual Review Pathology: Mechanisms Disease. 2021; 16. DOI: 10.1146/annurev-pathmechdis-012419-032626
3. Merril C, Geier M, and Petricciani J. Bacterial Virus Gene Expression in Human Cells. Nature 1971; 233. DOI: 10.1038/233398a0
4. Uddin F, Rudin CM, and Sen T. CRISPR Gene Therapy: Applications, Limitations, and Implications for the Future. Front Oncol. 2020; 10(1387). DOI: 10.3389/fonc.2020.01387
5. Wimberger S, Akrap N, Firth M, Brengdahl J, Engberg S, Schwinn MK, Slater MR, and Lundin A. Simultaneous inhibition of DNA-PK and Polϴ improves integration efficiency and precision of genome editing. Nat. Commun. 2023; 14. DOI: 10.1038/s41467-023-40344-4
6. Zhang XH, Tee LY, Wang XG, Huang QS, and Yang SH. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. Mol Ther Nucleic Acids. 2015; 4(11). DOI: 10.1038/mtna.2015.37
7. Cong L and Zhang F. Genome Engineering Using CRISPR-Cas9 System. *Chromosomal Mutagenesis*. Ed. by Pruett-Miller S. Humana Press, New York, NY, 2015. DOI: 10.1007/978-1-4939-1862-1_10
8. Heyer WD, Ehmsen KT, and Liu J. Regulation of homologous recombination in eukaryotes. Annu Rev Genet. 2010; 44. DOI: 10.1146/annurev-genet-051710-150955
9. Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, Chen PJ, Wilson C, Newby GA, Raguram A, and Liu DR. Search-and-replace genome editing without double-strand breaks or donor DNA. Nat. 2019; 576. DOI: 10.1038/s41586-019-1711-4
10. Mentani A, Maresca M, and Shiriaeva A. Prime Editing: Mechanistic Insights and DNA Repair Modulation. Cells 2025; 14(4). DOI: 10.3390/cells14040277
11. Doman JL, Pandey S, Neugebauer ME, Tolar J, Osborn MJ, and Liu DR. Phage-assisted evolution and protein engineering yield compact, efficient prime editors. Cell 2023; 186:3983–4002. DOI: 10.1016/j.cell.2023.07.039

12. Manghwar H, Lindsey K, Zhang X, and Jin S. CRISPR/Cas System: Recent Advances and Future Prospects for Genome Editing. Trends in Plant Science 2019; 24(12). DOI: 10.1016/j.tplants.2019.09.006
13. Khan SH. Genome-Editing Technologies: Concept, Pros, and Cons of Various Genome-Editing Techniques and Bioethical Concerns for Clinical Application. Molecular Therapies 2019; 16. DOI: https://doi.org/10.1016/j.omtn.2019.02.027
14. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, and Charpentier E. A programmable dual RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 2012; 28(337). DOI: 10.1126/science.1225829
15. Pourcel C, Salvignol G, and Vergnaud G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology 2005; 151(3). DOI: 10.1099/mic.0.27437-0
16. Wan L and Han H. Strategies for improving the genome-editing efficiency of class 2 CRISPR/Cas system. Cell Press 2024; 10(19). DOI: 10.1016/j.heliyon.2024.e38588
17. Ihry RJ, Worringer KA, Salick MR, Frias E, Ho D, Theriault K, Kommineni S, Chen J, Sondey M, Ye C, Randhawa R, Kulkarni T, Yang Z, McAllister G, Russ C, Reece-Hoyes J, Forrester W, Hoffman GR, Dolmetsch R, and Kaykas A. p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. Nat Med. 2018; 24(7). DOI: 10.1038/s41591-018-0050-6
18. Kosicki M, Tomberg K, and Bradley A. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. Nat Biotechnol. 2019; 36(8). DOI: 10.1038/nbt.4192
19. Eid A, Alshareef S, and Mahfouz MM. CRISPR base editors: genome editing without double-stranded breaks. Biochem J. 2018; 475(11). DOI: 10.1042/BCJ20170793
20. Nelson JW, Randolph PB, Shen SP, Everette KA, Chen PJ, Anzalone AV, An M, Newby GA, Chen JC, Hsu A, and Liu DR. Engineered pegRNAs improve prime editing efficiency. Nat Biotechnol 2022; 40. DOI: 10.1038/s41587-021-01039-7
21. Li X, Wang X, Sun W, Huang S, Zhong M, Yao Y, Ji Q, and Huang X. Enhancing prime editing efficiency by modified pegRNA with RNA G-quadruplexes. Journal of Molecular Cell Biology 2022; 14(4). DOI: 10.1093/jmcb/mjac022
22. Zhang G, Liu Y, and Huang S. Enhancement of prime editing via xrRNA motif-joined pegRNA. Nat Commun 2022; 13. DOI: 10.1038/s41467-022-29507-x
23. Liu Y, Yang G, Huang S, Li X, Wang X, Li G, Chi T, Chen Y, Huang X, and Wang X. Enhancing prime editing by Csy4-mediated processing of pegRNA. Cell Res 2021; 31. DOI: 10.1038/s41422-021-00520-x
24. Gao Z, Ravendran S, Mikkelsen NS, Haldrup J, Cai H, Ding X, Paludan SR, and K. TM. A truncated reverse transcriptase enhances prime editing by split AAV vectors. Mol. Ther. 2022; 30(9). DOI: 10.1016/j.ymthe.2022.07.001
25. Grünewald J, Miller BR, Szalay RN, Cabeceiras PK, Woodilla CJ, and Holtz EJB. Engineered CRISPR prime editors with compact, untethered reverse

transcriptases. Nat. Biotechnol. 2023; 41. DOI: 10.1038/s41587-022-01473-1

26. Harnessing DSB repair to promote efficient homology-dependent and -independent prime editing. Nat. Commun. 2022; 13. DOI: 10.1038/s41467-022-28771-1
27. Adikusuma F, Lushington C, Arudkumar J, Godahewa GI, Chey YCJ, Gierus L, and Piltz S. Optimized nickase- and nuclease-based prime editing in human and mouse cells. Nucleic Acids Res. 2021; 49(18). DOI: 10.1093/nar/gkab792
28. Li X, Zhang G, Huang S, Liu Y, Tang J, Zhong M, Wang X, and Sun W. Development of a versatile nuclease prime editor with upgraded precision. Nat. Commun. 2023; 14. DOI: 10.1038/s41467-023-35870-0
29. Kweon J, Yoon JK, Jang AH, Shin HR, See JE, Jang G, Kim JI, and Kim Y. Engineered prime editors with PAM flexibility. Molecular Therapy 2021; 9(6). DOI: 10.1016/j.ymthe.2021.02.022
30. Liang R, He Z, Zhao KT, Zhu H, Hu J, Liu G, Gao Q, Liu M, Zhang R, Qiu JL, and Gao C. Prime editing using CRISPR-Cas12a and circular RNAs in human cells. Nat Biotechnol 2024; 42. DOI: 10.1038/s41587-023-02095-x
31. Chen PJ, Hussmann JA, Yan J, Knipping F, Ravisankar P, Chen PF, and Chen C. Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. Cell 2021; 184(22). DOI: 10.1016/j.cell.2021.09.018
32. Ferreira da Silva J, Oliveira GP, Arasa-Verge EA, Kagiou C, Moretton A, Timelthaler G, Jiricny J, and Loizou JI. Prime editing efficiency and fidelity are enhanced in the absence of mismatch repair. Nat Commun 2022; 13(760). DOI: 10.1038/s41467-022-28442-1
33. Lee J, Kweon J, and Kim Y. Emerging trends in prime editing for precision genome editing. Exp Mol Med 2025; 57. DOI: 10.1038/s12276-025-01463-8
34. Doman JL, Sousa AA, Randolph PB, Chen PJ, and Liu DR. Designing and executing prime editing experiments in mammalian cells. Nat Protoc 2022; 17:2431–68. DOI: 10.1038/s41596-022-00724-4
35. Gvozdenovic A and Corn JE. Fueling next-generation genome editing with DNA repair. Curr Opin Biomed Eng 2023; 28. DOI: 10.1016/j.cobme.2023.100506
36. Huber LB, Betz K, and Marx A. Reverse Transcriptases: From Discovery and Applications to Xenobiology. ChemBioChem 2022; 24. DOI: 10.1002/cbic.202200521
37. Gladyshev EA and Arkhipova IR. A widespread class of reverse transcriptase-related cellular genes. Proc. Natl. Acad. Sci. U.S.A. 2011; 108(51). DOI: 10.1073/pnas.1100266108
38. Kelleher C, Teixeira MT, Förstemann K, and Lingner J. Telomerase: biochemical considerations for enzyme and substrate. Trends Biochem Sci. 2002; 27(11). DOI: 10.1016/s0968-0004(02)02206-5
39. Haendeler J, Dröse S, Büchner N, Jakob S, Altschmied J, Goy C, Spyridopoulos I, Zeiher AM, Brandt U, and Dimmeler S. Mitochondrial telomerase reverse transcriptase binds to and protects mitochondrial DNA and function from damage. Arterioscler Thromb Vasc Biol. 2009; 29(6). DOI: 10.1161/ATVBAHA.109.185546

40. Finnegan DJ. Eukaryotic transposable elements and genome evolution. Trends Genet. 1989; 5(4). DOI: 10.1016/0168-9525(89)90039-5
41. Martín-Alonso S, Frutos-Beltrán E, and Menéndez-Arias L. Reverse Transcriptase: From Transcriptomics to Genome Editing. Trends Biotecnol. 2021; 39(2). DOI: 10.1016/j.tibtech.2020.06.008
42. Oscorbin IP and Filipenko ML. M-MuLV reverse transcriptase: Selected properties and improved mutants. Comput Struct Biotechnol J. 2021; 19. DOI: 10.1016/j.csbj.2021.11.030
43. Das D and Georgiadis MM. The Crystal Structure of the Monomeric Reverse Transcriptase from Moloney Murine Leukemia Virus. Cell Press 2004; 12(5). DOI: 10.1016/j.str.2004.02.032
44. Doublié S, Sawaya MR, and Ellenberger T. An open and closed case for all polymerases. Structure 1999; 7(2). DOI: 10.1016/S0969-2126(99)80017-3
45. Coté ML and Roth MJ. Murine Leukemia Virus Reverse Transcriptase: Structural Comparison with HIV-1 Reverse Transcriptase. Virus Res. 2008; 134(1-2). DOI: 10.1016/j.virusres.2008.01.001
46. Schultz SJ and Champoux JJ. RNase H activity: structure, specificity, and function in reverse transcription. Virus research 2008; 134(1-2). DOI: 10.1016/j.virusres.2007.12.007
47. Hsiou Y, Ding J, Das K, Clark AJ, Hughes S, and Arnold E. Structure of unliganded HIV-1 reverse transcriptase at 2.7 A resolution: implications of conformational changes for polymerization and inhibition mechanisms. Structure 1996; 4(7). DOI: 10.1016/s0969-2126(96)00091-3
48. Johnson KA. Conformational coupling in DNA polymerase fidelity. Annu Rev Biochem. 1993; 62. DOI: 10.1146/annurev.bi.62.070193.003345.
49. Yang L, Arora K, Beard WA, Wilson SH, and Schlick T. Critical role of magnesium ions in DNA polymerase beta's closing and active site assembly. J Am Chem Soc. 2004; 126(27). DOI: 10.1021/ja049412o
50. Ralec C, Henry E, Lemor M, and Killelea Henneke TG. Calcium-driven DNA synthesis by a high-fidelity DNA polymerase. Nucleic Acids Res. 2017; 45(21). DOI: 10.1093/nar/gkx927
51. Hejna J and Moses R. DNA Polymerases. *Encyclopedia of Microbiology*. Ed. by Schaechter M. 3rd ed. ISBN: 978-0-12-373944-5. Academic Press, 2009
52. Yin X, Popa H, Stapon A, Bouda E, and Garcia-Diaz M. Fidelity of Ribonucleotide Incorporation by the SARS-CoV-2 Replication Complex. Journal of Molecular Biology 2023; 435(5). DOI: 10.1016/j.jmb.2023.167973
53. analytikjena. Thermal Shift Assay using SYPRO Orange to Detect Protein Melting Temperatures. Available from: https://www.analytik-jena.com/knowledge/applications/application-teasers/thermal-shift-assay-using-sypro-orange-to-detect-protein-melting-temperatures/ [Accessed on: 2026 Mar 2]
54. Singh A, Upadhyay V, Singh A, and Panda AK. Structure-Function Relationship of Inclusion Bodies of a Multimeric Protein. Sec. Microbial Physiology and Metabolism 2020; 11. DOI: 10.3389/fmicb.2020.00876

55. Idicula-Thomas S and Balaji PV. Correlation Between the Structural Stability and Aggregation Propensity of Proteins. In Silico Biol. 2007; 7(2). PMID: 17688448
56. Graham TGW, Dugast-Darzacq C, Dailey GM, Wrynla XH, Van Dis E, Esbin MN, Abidi A, Stanley SA, Darzacq X, and Tjian R. Open-source RNA extraction and RT-qPCR methods for SARS-CoV-2 detection. PLoS One 2021; 16(2). DOI: 10.1371/journal.pone.0246647
57. Scientific T. SuperScript IV Reverse Transcriptase. Available from: https://www.thermofisher.com/se/en/home/life-science/pcr/reverse-transcription/superscript-iv-reverse-transcriptase.html [Accessed on: 2026 Mar 3]
58. Wong A. The Ethics of HEK 293. 2006 :473–95
59. Trojan J, Zeuzem S, Randolph A, Hemmerle C, Brieger A, Raedle J, Plotz G, Jiricny J, and Marra G. Functional analysis of hMLH1 variants and HNPCC-related mutations using a human expression system. Gastroenterol 2002; 122(1). DOI: 10.1053/gast.2002.30296
60. Lin Y, Wagner E, and Lächelt U. Non-viral delivery of the CRISPR/Cas system: DNA versus RNA versus RNP. Biomaterials Science 2022; 8(5). DOI: 10.1039/D1BM01658J
61. Promega. Transfection Reagent Basics. Available from: https://se.promega.com/products/luciferase-assays/transfection-reagents/#related-resources [Accessed on: 2026 May 4]
62. Scientific T. CRISPR-Cas9 Protein for Genome Editing—Getting Started Guide. Available from: https://www.thermofisher.com/se/en/home/life-science/genome-editing/genome-editing-learning-center/genome-editing-resource-library/genome-editing-application-notes/crispr-cas9-protein-genome-editing-getting-started-guide.html [Accessed on: 2026 Mar 1]
63. Illumina. Accelerate CRISPR genome editing with next-generation sequencing. Available from: https://www.illumina.com/techniques/popular-applications/genome-editing.html#:~:text=More%20Uses%20of%20NGS%20in,experiments%20with%20large%20plasmid%20libraries. [Accessed on: 2026 Mar 1]
64. Sheffield BS, Beharry A, Diep J, and Perdrizet K. Point of Care Molecular Testing: Community-Based Rapid Next-Generation Sequencing to Support Cancer Care. Current Oncology 2022; 29(3). DOI: 10.3390/curroncol29030113
65. Lee Y, Clark EW, Milan MSD, Champagne C, Michael KS, Awad MM, Barbie DA, Cheng ML, Kehl KL, Marcoux JP, Rabin MS, Rotow JK, Sands JM, Jänne PA, and Oxnard GR. Turnaround Time of Plasma Next-Generation Sequencing in Thoracic Oncology Patients: A Quality Improvement Analysis. JCO Precis Oncol. 2020; 4. DOI: 10.1200/PO.20.00121
66. Lee J. HiBiT: A Tiny Tag Expanding Endogenous Protein Detection. Promega Biotech AB 2024
67. Connections P. Just What Is an RLU (Relative Light Unit)? Available from: https://www.promegaconnections.com/measuring-relative-light-units/ [Accessed on: 2026 May 5]

68. Dacquay LC, Choong OK, Membrino A, Lundin A, Akrap N, Kölnner J, Hsieh PP, Baran A, Gordon E, Mustfa S, Saka S, Crysnanto D, Karlsson F, Dahlen A, Thom G, Lindgren J, Firth M, Schwinn MK, Machleidt T, Peterka M, Svikovic S, Sienski G, and Maresca M. Artemis Regulates Homology-Independent Prime Editing (PRINS) for Enhanced Genomic Insertions. [preprint]. DOI: 10.64898/2025.12.08.693005. [Accessed on: 2026 Mar 2]
69. Atlas THP. GAPDH. Available from: https://www.proteinatlas.org/ENSG00000111640-GAPDH/cell+line [Accessed on: 2026 Feb 12]
70. Atlas THP. TUBA1A. Available from: https://www.proteinatlas.org/ENSG00000167552-TUBA1A [Accessed on: 2026 Mar 14]
71. Atlas THP. ACTB. Available from: https://www.proteinatlas.org/ENSG00000075624-ACTB [Accessed on: 2026 Mar 14]
72. Atlas THP. EEF1A1. Available from: https://www.proteinatlas.org/ENSG00000156508-EEF1A1/cell+line [Accessed on: 2026 Mar 14]
73. Toktay Y, Dayanc B, and Senturk S. Engineering and validation of a dual luciferase reporter system for quantitative and systematic assessment of regulatory sequences in Chinese hamster ovary cells. Sci. Rep. 2022; 12(1). DOI: 10.1038/s41598-022-09887-2
74. Qin JY, Zhang L, Clift KL, Hulur I, Xiang AP, Ren BZ, and Lahn BT. Systematic Comparison of Constitutive Promoters and the Doxycycline-Inducible Promoter. PLoS One. 2010; 5(5). DOI: 10.1371/journal.pone.0010611
75. Tornøe J, Kusk P, Johansen TE, and Jensen PR. Generation of a synthetic mammalian promoter library by modification of sequences spacing transcription factor binding sites. Gene 2002; 297(1-2). DOI: 10.1016/S0378-1119(02)00878-8
76. Jia Y, Li S, Allen G, Feng S, and Xue L. A Novel Glyceraldehyde-3-Phosphate Dehydrogenase (GAPDH) Promoter for Expressing Transgenes in the Halotolerant Alga Dunaliella salina. Curr Microbiol. 2012; 64. DOI: 10.1007/s00284-012-0102-y
77. Quinn CM, Wiles AP, El-Shanawany T, Catchpole I, Alnadaf T, Ford MJ, Gordon S, and Greaves DR. The human eukaryotic initiation factor 4AI gene (EIF4A1) contains multiple regulatory elements that direct high-level reporter gene expression in mammalian cell lines. Genomics 1999; 62(3). DOI: 10.1006/geno.1999.6031
78. Herweijer H, Zhang G, Subbotin VM, Budker V, Williams P, and Wolff JA. Time course of gene expression after plasmid DNA gene transfer to the liver. The Journal of Gene Medicine 2001; 3(3). DOI: 10.1002/jgm.178
79. Gazdhar A, Bilici M, Pierog J, Ayuni EL, Gugger M, Wetterwald A, Cecchini M, and Schmid RA. In vivo electroporation and ubiquitin promoter – a protocol for sustained gene expression in the lung. The Journal of Gene Medicine 2006; 8(7). DOI: 10.1002/jgm.911
80. García-Olloqui P, Rodriguez-Madoz J, Di Scala M, Abizanda G, Vales A, Olagüe C, Iglesias O, Larequi E, Aguado Alvaro L, Ruiz A, Prosper F, Gonzalez-Aseguinolaza G, and Pelacho B. Effect of heart ischemia and administration route on biodistribution and transduction efficiency of AAV9 vectors. Journal

of Tissue Engineering and Regenerative Medicine 2019; 14. DOI: 10.1002/term.2974

81. Norrman K, Fischer Y, Bonnamy B, Wolfhagen Sand F, Ravassard P, and Semb H. Quantitative Comparison of Constitutive Promoters in Human ES cells. PLoS One 2010; 5(8). DOI: 10.1371/journal.pone.0012413
82. D'Aiuto L, Robison C, Gigante M, Nwanegbo E, Shaffer B, Sukhwani M, Castro C, and Chaillet J. Human IL-12 p40 as a reporter gene for high-throughput screening of engineered mouse embryonic stem cells. BMC biotechnology 2008; 8:52. DOI: 10.1186/1472-6750-8-52
83. Kwak CS, Oflaz FE, Qiu J, and Wang X. Human stem cell-specific epigenetic signatures control transgene expression. Biochim Biophys Acta Gene Regul Mech. 2024; 1867(4). DOI: 10.1016/j.bbagrm.2024.195063
84. Instruments E. What is Fluorescence Anisotropy? 2024. Available from: https://www.edinst.com/resource/what-is-fluorescence-anisotropy/ [Accessed on: 2025 Jun 26]
85. GraphPad. Equation: Specific binding with Hill slope. 2025. Available from: https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_specific_hill.htm [Accessed on: 2025 Sep 25]
86. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, and Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature 2021; 596. DOI: 10.1038/s41586-021-03819-2
87. Passaro S, Corso G, Wohlwend J, Reveiz M, Thaler Vignesh S, Somnath R, Getz N, Portnoi T, Roy J, Stark H, Kwabi-Addo D, Beaini D, Jaakkola T, and Barzilay R. Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. 2025; 596. [preprint]. DOI: 10.1101/2025.06.14.659707. [Accessed on: 2026 Mar 2]
88. Schrödinger. PyMOL3. Available from: https://www.schrodinger.com/platform/products/pymol/ [Accessed on: 2026 Mar 2]
89. Wiki P. Practical Pymol for Beginners. Available from: https://pymolwiki.org/index.php/Practical_Pymol_for_Beginners [Accessed on: 2026 Mar 2]
90. Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, Wicky BIM, Courbet A, Haas RJ de, Bethel N, Leung PJY, Huddy TF, Pellock S, Tischer D, Chan F, Koepnick B, Nguyen H, Kang A, Sankaran B, Bera AK, King NP, and Baker D. Robust deep learning-based protein sequence design using ProteinMPNN. Science 2022; 378(6615). DOI: 10.1126/science.add2187
91. Benjaminson E. Understanding Message Passing in GNNs. 2022 Jun 19. Available from: https://sassafras13.github.io/GNN/ [Accessed on: 2026 Mar 2]
92. BioLM. Soluble ProteinMPNN. Available from: https://biolm.ai/models/soluble-mpnn/ [Accessed on: 2026 Mar 22]

93. Dieckhaus H, Brocidiacono M, Randolph NZ, and Kuhlman B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. Proc. Natl. Acad. Sci. U.S.A. 2024; 121(6). DOI: 10.1073/pnas.2314853121
94. Dieckhaus H and Kuhlman B. Protein stability models fail to capture epistatic interactions of double point mutations. 2024. [preprint]. DOI: 10.1101/2024.08.20.608844. [Accessed on: 2026 Mar 2]
95. QIAGEN. Quick-Start Protocol: QIAGEN® Plasmid Plus Maxi Kit. Identifier: HB-0595-002. 2015 Oct. Available from: https://www.qiagen.com/en-US/resources/download/Protocols/qiagen-plasmid-plus-maxi-kit-en
96. Scientific T. High Capacity cDNA Reverse Transcription Kit USER GUIDE. Publication number: MAN0017977. Available from: https://documents.thermofisher.com/TFS-Assets/LSG/manuals/MAN0017977_highcap_cDNA_RT_UG.pdf [Accessed on: 2026 Mar 16]
97. Scientific T. Pierce BCA Protein Assay Kit USER GUIDE. Publication number: MAN0011430. Available from: https://documents.thermofisher.com/TFS-Assets/LSG/manuals/MAN0011430_Pierce_BCA_Protein_Asy_UG.pdf [Accessed on: 2026 Mar 16]
98. Laboratories BR. General Protocol for Western Blotting. Available from: https://www.bio-rad.com/webroot/web/pdf/lsr/literature/Bulletin_6376.pdf
99. Laboratories BR. Trans-Blot Turbo Transfer System Transfer Pack Quick Start Guide
100. Mathews DH. RNA Secondary Structure Analysis Using RNAstructure. Curr Protoc Bioinformatics. 2006; (12.6). DOI: 10.1002/0471250953.bi1206s13
101. Sumida KH, Núñez-Franco R, Kalvet I, Pellock SJ, Wicky BIM, Milles LF, Dauparas J, Wang J, Kipnis Y, Jameson N, Kang A, De La Cruz J, Sankaran B, Bera AK, Jiménez-Osés G, and Baker D. Improving Protein Expression, Stability, and Function with ProteinMPNN. Journal of the American Chemical Society 2024; 146(3). DOI: DOI:10.1021/jacs.3c10941
102. Romani AM. Intracellular magnesium homeostasis. *Magnesium in the Central Nervous System [Internet]*. Ed. by Vink R and Nechifor M. ISBN-13: 9780987073051. Adelaide (AU): University of Adelaide Press, 2011
103. Chandramouly G, Zhao J, McDevitt S, Rusanov T, Hoang T, Borisonnik N, Treddinick T, Lopezcolorado FW, and Kent T. Pol reverse transcribes RNA and promotes RNA-templated DNA repair. Sci. Adv. 2021; 7(24). DOI: 10.1126/sciadv.abf1771

# A

# Sequences

## A.1 Protein Sequences

This section features all the protein sequences of the panel of reference RTs.

**>pol11ERV_RT**
ATVEPPKPIPLTWKTEKPVWVNQWPLPKQKLEALHLLANEQLEKGHIEPS
FSPWNSPVFVIQKKSGKWRMLTDLRAVNAVIQPMGPLQPGLPSPAMIPKD
WPLIIIDLKDCFFTIPLAEQDCEKFAFTIPAINNKEPATRFQWKVLPQGMLN
SPTICQTFVGRALQPVREKFSDCYIIHYIDDILCAAETKDKLIDCYTFLQAEV
ANAGLAIASDKIQTSTPFHYLGMQIENRKIKPQKIEIRKDTLKTLNDFQKLLG
DINWIRPTLGIPTYAMSNLFSILRGDSDLNSKRILTPEATKEIKLVEEKIQSAQ
INRIDPLAPLQLLIFATAHSPTGIIIQNTDLVEWSFLPHSTVKTFTLYLDQIAT
LIGQTRLRIIKLCGNDPDKIVVPLTKEQVRQAFINSGAWQIGLANFVGIIDNH
YPKTKIFQFLKMTTWILPKITRREPLENALTVFTDGSSNGKAAYTGPKERV
IKTPYQSAQRAELVAVITVLQDFDQPINIISDSAYVVQATRDVETALIKYSMD
DQLNQLFNLLQQTVRKRNFPFYITHIRAHTNLPGPLTKANEEADLLVS

**>wtMMLV**
TLNIEDEYRLHETSKEPDVSLGSTWLSDFPQAWAETGGMGLAVRQAPLIIP
LKATSTPVSIKQYPMSQEARLGIKPHIQRLLDQGILVPCQSPWNTPLLPVKK
PGTNDYRPVQDLREVNKRVEDIHPTVPNPYNLLSGLPPSHQWYTVLDLKD
AFFCLRLHPTSQPLFAFEWRDPEMGISGQLTWTRLPQGFKNSPTLFDEALH
RDLADFRIQHPDLILLQYVDDLLLAATSELDCQQGTRALLQTLGNLGYRASA
KKAQICQKQVKYLGYLLKEGQRWLTEARKETVMGQPTPKTPRQLREFLG
TAGFCRLWIPGFAEMAAPLYPLTKTGTLFNWGPDQQKAYQEIKQALLTAP
ALGLPDLTKPFELFVDEKQGYAKGVLTQKLGPWRRPVAYLSKKLDPVAAG
WPPCLRMVAAIAVLTKDAGKLTMGQPLVILAPHAVEALVKQPPDRWLSNA
RMTHYQALLLDTDRVQFGPVVALNPATLLPLPEEGLQHNCLDILAEAHGTR
PDLTDQPLPDADHTWYTDGSSLLQEGQRKAGAAVTTETEVIWAKALPAGT
SAQRAELIALTQALKMAEGKKLNVYTDSRYAFATAHIHGEIYRRRGLLTSE
GKEIKNKDEILALLKALFLPKRLSIIHCPGHQKGHSAEARGNRMADQAARK
AAITETPDTSTLLIENSSP

**>MMLV_PE2**
TLNIEDEYRLHETSKEPDVSLGSTWLSDFPQAWAETGGMGLAVRQAPLIIPL
KATSTPVSIKQYPMSQEARLGIKPHIQRLLDQGILVPCQSPWNTPLLPVKKP
GTNDYRPVQDLREVNKRVEDIHPTVPNPYNLLSGLPPSHQWYTVLDLKDA
FFCLRLHPTSQPLFAFEWRDPEMGISGQLTWTRLPQGFKNSPTLFNEALHR
DLADFRIQHPDLILLQYVDDLLLAATSELDCQQGTRALLQTLGNLGYRASAK
KAQICQKQVKYLGYLLKEGQRWLTEARKETVMGQPTPKTPRQLREFLGK
AGFCRLFIPGFAEMAAPLYPLTKPGTLFNWGPDQQKAYQEIKQALLTAPAL
GLPDLTKPFELFVDEKQGYAKGVLTQKLGPWRRPVAYLSKKLDPVAAGW
PPCLRMVAAIAVLTKDAGKLTMGQPLVILAPHAVEALVKQPPDRWLSNAR
MTHYQALLLDTDRVQFGPVVALNPATLLPLPEEGLQHNCLDILAEAHGTRP
DLTDQPLPDADHTWYTDGSSLLQEGQRKAGAAVTTETEVIWAKALPAGTS
AQRAELIALTQALKMAEGKKLNVYTDSRYAFATAHIHGEIYRRRGWLTSEG
KEIKNKDEILALLKALFLPKRLSIIHCPGHQKGHSAEARGNRMADQAARKAA
ITETPDTSTLLIENSSP

**>MMLV_PE6d**
TLNIEDEYRLHETSKEPDVSLGSTWLSDFPQAWAETGGMGLAVRQAPLIIPL
KATSTPVSIKQYPMSQEARLGIKPHIQRLLDQGILVPCQSPWNTPLLPVKKP
GTNDYRPVQDLREVNKRVEDIHPNVPNPYNLLSGLPPSHQWYTVLDLKDA
FFCLRLHPTSQPLFAFEWRDPEMGISGQLTWTRLPQGFKNSPTLFCEALHR
DLADFRIQHPDLILLQYYDDLLLAATSELDCQQGTRALLQTLGNLGYRASAK
KAQICQKQVKYLGYLLKEGQRWLTEARKETVMGQPTPKTPRQLREFLGK
AGFCRLFIPGFAEMAAPLYPLTKPGTLFNWGPDQQKAYQEIKQALLTAPAL
GLPDLTKPFELFVDEKQGYAKGVLTQKLGPWRRPVAYLSKKLDPVAAGW
PPCLRMVAAIAVLTKDAGKLTMGQPLVILAPHAVEALVKQPPDRWLSNAR
MTHYQALLLDTDRVQFGPVVALNPATLLPLPEEGLQHNCLD

**>MMLVΔRNaseH**
TLNIEDEYRLHETSKEPDVSLGSTWLSDFPQAWAETGGMGLAVRQAPLIIPL
KATSTPVSIKQYPMSQEARLGIKPHIQRLLDQGILVPCQSPWNTPLLPVKKP
GTNDYRPVQDLREVNKRVEDIHPTVPNPYNLLSGLPPSHQWYTVLDLKDA
FFCLRLHPTSQPLFAFEWRDPEMGISGQLTWTRLPQGFKNSPTLFNEALHR
DLADFRIQHPDLILLQYVDDLLLAATSELDCQQGTRALLQTLGNLGYRASAK
KAQICQKQVKYLGYLLKEGQRWLTEARKETVMGQPTPKTPRQLREFLGK
AGFCRLFIPGFAEMAAPLYPLTKPGTLFNWGPDQQKAYQEIKQALLTAPAL
GLPDLTKPFELFVDEKQGYAKGVLTQKLGPWRRPVAYLSKKLDPVAAGW
PPCLRMVAAIAVLTKDAGKLTMGQPLVILAPHAVEALVKQPPDRWLSNAR
MTHYQALLLDTDRVQFGPVVALNPATLLPLPEEGLQHNCLD

**>MMLV_MK**
TLNIEDEYRLHETSKEPDVSLGSTWLSDFPQAWAETGGMGLAVRQAPLIIPL
KATSTPVSIKQYPMSQEARLGIKPHIQRLLDQGILVPCQSPWNTPLLPVKKP
GTNDYRPVQDLREVNKRVEDIHPTVPNPYNLLSGLPPSHQWYTVLDLKDA
FFCLRLHPTSQPLFAFEWRDPEMGISGQLTWTRLPQGFKNSPTLFNEALHR
DLADFRIQHPDLILLQYMDDLLLAATSELDCQQGTRALLQTLGNLGYRASA

KKAQICQKQVKYLGYLLKEGQRWLTEARKETVMGQPTPKTPRQLREFLG
KAGFCRLFIPGFAEMAAPLYPLTKPGTLFNWGPDQQKAYQEIKQALLTAP
ALGLPDLTKPFELFVDEKQGYAKGVLTQKLGPWRRPVAYLSKKLDPVAAG
WPPCLRMVAAIAVLTKDAGKLTMGQPLVIKAPHAVEALVKQPPDRWLSNA
RMTHYQALLLDTDRVQFGPVVALNPATLLPLPEEGLQHNCLDILAEAHGTR
PDLTDQPLPDADHTWYTDGSSLLQEGQRKAGAAVTTETEVIWAKALPAGT
SAQRAELIALTQALKMAEGKKLNVYTDSRYAFATAHIHGEIYRRRGWLTSE
GKEIKNKDEILALLKALFLPKRLSIIHCPGHQKGHSAEARGNRMADQAARK
AAITETPDTSTLLIENSSP

**>bear_MMLV**
TLNIEDEHRLHETSKEPDVSLGSTWLSDFPQAWAETGGMGLAVRQAPLIIPL
KATSTPVSIKQYPMSQEARLGIKPHIQRLLDQGILVPCQSPWNTPLLPVKKP
GTNDYRPVQDLREVNKRVEDIHPTVPNPYNLLSGLPPSHQWYTVLDLKDA
FFCLRLHPTSQPLFAFEWRDPEMGISGQLTWTRLPQGFKNSPTLFDEALRR
DLADFRIQHPDLILLQYVDDLLLAATSELDCQQGTRALLQTLGNLGYRASAK
KAQICQKQVKYLGYLLKEGQRWLTEARKETVLGQPTPKTPRQLREFLGKA
GNCRLWIPGFAEMAAPLYPLTKTGTLFNWGPDQQKAYQEIKQALLTAPAL
GLPDLTKPFELFVDEKQGYAKGVLTQKLGPWRRPVAYLSKKLDPVAAGW
PPCLRMVAAIAVLTKDAGKLTMGQPLVILAPHAVEALVKQPPDRWLSNAR
MTHYQALLLDTDRVQFGPVVALNPATLLPLPEEGLQHNCLDILAEAHGTRP
DLTDQPLPDADHTWYTGGSSLLQEGQRKAGAAVTTETEVIWAKALPAGTS
AQRAQLIALTQALKMAEGKKLNVYTNSRYAFATAHIHGEIYRRRGLLTSEG
KEIKNKDEILALLKALFLPKRLSIIHCPGHQKGHSAEARGNRMADQAARKAA

**>wtTf1**
SISSSKHTLSQMNKVSNIVKEPELPDIYKEFKDITADTNTEKLPKPIKGLEFE
VELTQENYRLPIRNYPLPPGKMQAMNDEINQGLKSGIIRESKAINACPVMFV
PKKEGTLRMVVDYKPLNKYVKPNIYPLPLIEQLLAKIQGSTIFTKLDLKSAY
HLIRVRKGDEHKLAFRCPRGVFEYLVMPYGISTAPAHFQYFINTILGEAKES
HVVCYMDDILIHSKSESEHVKHVKDVLQKLKNANLIINQAKCEFHQSQVKFI
GYHISEKGFTPCQENIDKVLQWKQPKNRKELRQFLGSVNYLRKFIPKTSQL
THPLNKLLKKDVRWKWTPTQTQAIENIKQCLVSPPVLRHFDFSKKILLETD
ASDVAVGAVLSQKHDDDKYYPVGYYSAKMSKAQLNYSVSDKEMLAIIKSLK
HWRHYLESTIEPFKILTDHRNLIGRITNESEPENKRLARWQLFLQDFNFEINY
RPGSANHIADALSRIVDETEPIPKDSEDNSINFVNQISIS

**>Tf1_PE6b**
SISSSKHTLSQMNKVSNIVKEPELPDIYKEFKDITADTNTEKLPKPIKGLEFE
VELTQENYRLPIRNYPLTPVKMQAMNDEINQGLKGGIIRESKAINACPVIFV
PRKEGTLRMVVDYRPLNKYVKPNVYPLPLIEQLLAKIQGSTIFTKLDLKSAY
HQIRVRKGDEHKLAFRCPRGVFEYLVMPYGISTAPAHFQYFINTILGEAKES
HVVCYMDDILIHSKSESEHVKHVKDVLQKLKNANLIINQAKCEFHQSQVKFI
GYHISEKGLTPCQENIDKVLQWKQPKNRKELRQFLGSVNYLRKFIPKTSQL
THPLNKLLKKDVRWKWTPTQTQAIENIKQCLVSPPVLRHFDFSKKILLETD
VSDVAVGAVLSQKHDDDKYYPVGYYSAKMSKAQLNYSVSDKEMLAIIKSLE

HWRHYLESTIEPFKILTDHRNLIGRITNESEPENKRLARWQLFLQDFNFEINY
RPGSANHIADALSRIVDETEPIPKDNEDNSINFVNQISIS

**>Tf1_PE6c**
SISSSKHTLSQMNKVSNIVKEPELPDIYKEFKDITADTNTEKLPKPIKGLEFE
VELTQENYRLPIRNYPLTPVKMQAMNDEINQGLKGGIIRESKAINACPVIFV
PRKEGTLRMVVDYRPLNKYVKPNVYPLPLIEQLLAKIQGSTIFTKLDLKSAY
HQIRVRKGDEHKLAFRCPRGVFEYLVMPYGIKTAPAHFQYFINTILGEAKE
SHVVCYMDDILIHSKSESEHVKHVKDVLQKLKNANLIINQAKCEFHQSQVKF
LGYHISEKGLTPCQENIDKVLQWKQPKNQKELRQFLGQVNYLRKFIPKTSQ
LTHPLNKLLKKDVRWKWTPTQTQAIENIKQCLVSPPVLRHFDFSKKILLET
DVSDVAVGAVLSQKHDDDKYYPVGYYSAKMSKAQLNYSVSDKEMLAIIKSL
EHWRHYLESTIEPFKILTDHRNLIGRITNESEPENKRLARWQLFLQDFNFEIN
YRPGSANHIADALSRIVDETEPIPKDNEDNSINFVNQISIS

**>wtEc48**
SGRPYVTLNLNGMFMDKFKPYSKSNAPITTLEKLSKALSISVEELKAIAELSL
DEKYTLKEIPKIDGSKRIVYSLHPKMRLLQSRINKRIFKELVVFPSFLFGSVPS
KNDVLNSNVKRDYVSCAKAHCGAKTVLKVDISNFFDNIHRDLVRSVFEEILH
IKDEALEYLVDICTKDDFVVQGALTSSYIATLCLFAVEGDVVRRAQRKGLV
YTRLVDDITVSSKISNYDFSQMQSHIERMLSEHDLPINKHKTKIFHCSSEPIKV
HGLRVDYDSPRLPSDEVKRIRASIHNLKLLAAKNNTKTSVAYRKEFNRCMG
RVNKLGRVGHEKYESFKKQLQAIKPMPSKRDVAVIDAAIKSLELSYSKGNQ
NKHWYKRKYDLTRYKMIILTRSESFKEKLECFKSRLASLKPLS

**>Ec48 PE6a**
GRPYVTLNLNGMFMDKFKPYSKSNAPITTLEKLSKALSISVEELKAIAELSLD
EKYTLKKIPKIDGSKRIVYSLHPKMRLLQSRINERIFKELVVFPSFLFGSVPSK
NDVLNSNVKRDYVSCAKAHCGAKTVLKVDISNFFDNIHRDLVRSVFEEILHI
KDEALDYLVDICTKDDFVVQGALTSSYIATLCLFAVEGDVVRRAQRKGLVY
TRLVDDITVSSKISNYDFSQMQSHIERMLSEHNLPINKHKTKIFHCSSEPIKVH
GLIVDYDSPRLPSDKVKRIRASIHNLKLLAAKNNTKTSVAYRKEFNRCMGR
VNELGRVGHEKYESFKKQLQAIKPMPSNRDVAVIDAAIKSLELSYSKGNQN
KHWYKRKYDLTRYKMIILTRSESFKEKLECFKSRLASLKPL

## A.2 Promoter Sequences

This section shows all promoter sequences used.

**>CMV**
GACATTGATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCAT
TAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAA
ATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAAT
AATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACG
TCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCA
AGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAA

ATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTTCCT
ACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATG
CGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGG
GGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGG
CACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCAT
TGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCA
GAGCT

**>EFS**
GCTCCGGTGCCCGTCAGTGGGCAGAGCGCACATCGCCCACAGTCCCCG
AGAAGTTGGGGGGAGGGGTCGGCAATTGATCCGGTGCCTAGAGAAGG
TGGCGCGGGGTAAACTGGGAAAGTGATGTCGTGTACTGGCTCCGCCTT
TTTCCCGAGGGTGGGGGAGAACCGTATATAAGTGCAGTAGTCGCCGTG
AACGTTCTTTTTCGCAACGGGTTTGCCGCCAGAACACAGG

**>EIF4A1**
CCTGCAGGGCCCACTAGTTGATTTCCTTCATCCCTGGCACACGTCCAGG
CAGTGTCGAATCCATCTCTGCTACAGGGGAAAACAAATAACATTTGAG
TCCAGTGGAGACCGGGAGCAGAAGTAAAGGGAAGTGATAACCCCCAGA
GCCCGGAAGCCTCTGGAGGCTGAGACCTCGCCCCCCTTGCGTGATAGG
GCCTACGGAGCCACATGACCAAGGCACTGTCGCCTCCGCACGTGTGAG
AGTGCAGGGCCCCAAGATGGCTGCCAGGCCTCGAGGCCTGACTCTTCT
ATGTCACTTCCGTACCGGCGAGAAAGGCGGGCCCTCCAGCCAATGAGG
CTGCGGGGCGGGCCTTCACCTTGATAGGCACTCGAGTTATCCAATGGT
GCCTGCGGGCCGGAGCGACTAGGAACTAACGTCATGCCGAGTTGCTGA
GCGCCGGCAGGCGGGGCCGGGGCGGCCAAACCAATGCGATGGCCGGG
GCGGAGTCGGGCGCTCTATAAGTTGTCGATAGGCGGGCACTCCGCCCT
AGTTTCTAAGGA

**>GAPDH**
CCTGCAGGGCCCACTAGTTCCCCAACTTTCCCGCCTCTCAGCCTTTGAA
AGAAAGAAAGGGGAGGGGGCAGGCCGCGTGCAGTCGCGAGCGGTGCT
GGGCTCCGGCTCCAATTCCCCATCTCAGTCGCTCCCAAAGTCCTTCTGT
TTCATCCAAGCGTGTAAGGGTCCCCGTCCTTGACTCCCTAGTGTCCTGC
TGCCCACAGTCCAGTCCTGGGAACCAGCACCGATCACCTCCCATCGGG
CCAATCTCAGTCCCTTCCCCCCTACGTCGGGGCCCACACGCTCGGTGCG
TGCCCAGTTGAACCAGGCGGCTGCGGAAAAAAAAAAGCGGGGAGAAAG
TAGGGCCCGGCTACTAGCGGTTTTACGGGCGCACGTAGCTCAGGCCTC
AAGACCTTGGGCTGGGACTGGCTGAGCCTGGCGGGAGGCGGGGTCCGA
GTCACCGCCTGCCGCCGCGCCCCCGGTTTCTATAAATTGAGCCCGCAG
CCTCCCGCTTCGCTCTCTGCTCCTCCTGTTCGACAGTCAGCCGCATCTT
CTTTTGCGTCGCCAGGTGAAGACGGGCGGAGAGAAACCCGGGAGGCTA
GGGACGGCCTGAAGGCGGCAGGGGCGGGCGCAGGCCGGATGTGTTCG
CGCCGCTGCGGGGTGGGCCCGGGCGGCCTCCGCATTGCAGGGGCGGGC
GGAGGACGTGATGCGGCGCGGGCTGGGCATGGAGGCCTGGTGGGGGA
GGGGAGGGGAGGCGTGGGTGTCGGCCGGGGCCACTAGGCGCTCACTG

TTCTCTCCCTCCGCGCAGCCGAGCCACATCGCTGAGACAC

**>HSV_TK**
AAATGAGTCTTCGGACCTCGCGGGGGCCGCTTAAGCGGTGGTTAGGGT
TTGTCTGACGCGGGGGGAGGGGGAAGGAACGAAACACTCTCATTCGGA
GGCGGCTCGGGGTTTGGTCTTGGTGGCCACGGGCACGCAGAAGAGCGC
CGCGATCCTCTTAAGCACCCCCCCGCCCTCCGTGGAGGCGGGGGTTTG
GTCGGCGGGTGGTAACTGGCGGGCCGCTGACTCGGGCGGGTCGCGCGC
CCCAGAGTGTGACCTTTTCGGTCTGCTCGCAGACCCCCGGGCGGCGCC
GCCGCGGCGGCGACGGGCTCGCTGGGTCCTAGGCTCCATGGGGACCGT
ATACGTGGACAGGCTCTGGAGCATCCGCACGACTGCGGTGATATTACC
GGAGACCTTCTGCGGGACGAGCCGGGTCACGCGGCTGACGCGGAGCGT
CCGTTGGGCGACAAACACCAGGACGGGGCACAGGTACACTATCTTGTC
ACCCGGAGGCGCGAGGGACTGCAGGAGCTTCAGGGAGTGGCGCAGCTG
CTTCATCCCCGTGGCCCGTTGCTCGCGTTTGCTGGCGGTGTCCCCGGA
AGAAATATATTTGCATGTCTTTAGTTCTATGATGACACAAACCCCGCCC
AGCGTCTTGTCATTGGCGAAGTCGAACACGCAGATGCAGTCGGGGCGG
CGCGGTCCCAGGTCCACTTCGCATATTAAGGTGACGCGTGTGGCCTCG
AACACCGAGCGACCCTGCAGCGACCCGCTTAAAAGCTTGGCAATCCGG
TACT

**>JeT**
GAATTCGGGCGGAGTTAGGGCGGAGCCAATCAGCGTGCGCCGTTCCGA
AAGTTGCCTTTTATGGCTGGGCGGAGAATGGGCGGTGAACGCCGATGA
TTATATAAGGACGCGCCGGGTGTGGCACAGCTAGTTCCGTCGCAGCCG
GGATTTGGGTCGCGGTTCTTGTTTGTGGATCCCTGTGATCGTCACTTG
ACA

**>PGK**
GGGTAGGGGAGGCGCTTTTCCCAAGGCAGTCTGGAGCATGCGCTTTAG
CAGCCCCGCTGGGCACTTGGCGCTACACAAGTGGCCTCTGGCCTCGCA
CACATTCCACATCCACCGGTAGGCGCCAACCGGCTCCGTTCTTTGGTG
GCCCCTTCGCGCCACCTTCTACTCCTCCCCTAGTCAGGAAGTTCCCCCC
CGCCCCGCAGCTCGCGTCGTGCAGGACGTGACAAATGGAAGTAGCACG
TCTCACTAGTCTCGTGCAGATGGACAGCACCGCTGAGCAATGGAAGCG
GGTAGGCCTTTGGGGCAGCGGCCAATAGCAGCTTTGCTCCTTCGCTTT
CTGGGCTCAGAGGCTGGGAAGGGGTGGGTCCGGGGGCGGGCTCAGGG
GCGGGCTCAGGGGCGGGGCGGGCGCCCGAAGGTCCTCCGGAGGCCCG
GCATTCTGCACGCTTCAAAAGCGCACGTCTGCCGCGCTGTTCTCCTCTT
CCTCATCTCCGGGCCTTTCG

**>SV40**
CTGTGGAATGTGTGTCAGTTAGGGTGTGGAAAGTCCCCAGGCTCCCCA
GCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCAGG
TGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATG
CATCTCAATTAGTCAGCAACCATAGTCCCGCCCCTAACTCCGCCCATCC

CGCCCCTAACTCCGCCCAGTTCCGCCCATTCTCCGCCCCATGGCTGACT
AATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCTGCCTCTGAGCTA
TTCCAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAA
AAGCT

**>UbC**
GGTGCAGCGGCCTCCGCGCCGGGTTTTGGCGCCTCCCGCGGGCGCCCC
CCTCCTCACGGCGAGCGCTGCCACGTCAGACGAAGGGCGCAGCGAGCG
TCCTGATCCTTCCGCCCGGACGCTCAGGACAGCGGCCCGCTGCTCATA
AGACTCGGCCTTAGAACCCCAGTATCAGCAGAAGGACATTTTAGGACG
GGACTTGGGTGACTCTAGGGCACTGGTTTTCTTTCCAGAGAGCGGAAC
AGGCGAGGAAAAGTAGTCCCTTCTCGGCGATTCTGCGGAGGGATCTCC
GTGGGGCGGTGAACGCCGATGATTATATAAGGACGCGCCGGGTGTGGC
ACAGCTAGTTCCGTCGCAGCCGGGATTTGGGTCGCGGTTCTTGTTTGT
GGATCGCTGTGATCGTCACTTGGTGAGTAGCGGGCTGCTGGGCTGGCC
GGGGCTTTCGTGGCCGCCGGGCCGCTCGGTGGGACGGAAGCGTGTGGA
GAGACCGCCAAGGGCTGTAGTCTGGGTCCGCGAGCAAGGTTGCCCTGA
ACTGGGGGTTGGGGGGAGCGCAGCAAAATGGCGGCTGTTCCCGAGTCT
TGAATGGAAGACGCTTGTGAGGCGGGCTGTGAGGTCGTTGAAACAAGG
TGGGGGGCATGGTGGGCGGCAAGAACCCAAGGTCTTGAGGCCTTCGCT
AATGCGGGAAAGCTCTTATTCGGGTGAGATGGGCTGGGGCACCATCTG
GGGACCCTGACGTGAAGTTTGTCACTGACTGGAGAACTCGGTTTGTCG
TCTGTTGCGGGGGCGGCAGTTATGGCGGTGCCGTTGGGCAGTGCACCC
GTACCTTTGGGAGCGCGCGCCCTCGTCGTGTCGTGACGTCACCCGTTC
TGTTGGCTTATAATGCAGGGTGGGGCCACCTGCCGGTAGGTGTGCGGT
AGGCTTTTCTCCGTCGCAGGACGCAGGGTTCGGGCCTAGGGTAGGCTC
TCCTGAATCGACAGGCGCCGGACCTCTGGTGAGGGGAGGGATAAGTGA
GGCGTCAGTTTCTTTGGTCGGTTTTATGTACCTATCTTCTTAAGTAGCT
GAAGCTCCGGTTTTGAACTATGCGCTCGGGGTTGGCGAGTGTGTTTTG
TGAAGTTTTTTAGGCACCTTTTGAAATGTAATCATTTGGGTCAATATG
TAATTTTCAGTGTTAGACTAGTAAA

## A.3 Oligos

This section shows the main sequences of oligos used throughout the project, both DNA and RNA.

| **Oligo Name** | **Sequence** |
|---|---|
| GAPDH_DNA_primer | 5'-CATGGCCCACATGGCCT |
| GAPDH_RNA_template_blunt | 5'-CUCACCUCCUUGGAGGCCAUGUG GGCCAUG |
| GAPDH_RNA_template_24b | 5'-CUCACCUCCUUGGAGGCCAUGUGG |
| GAPDH_RNA_template_54b | 5'-UAGCUAAUCUUCUUGAACAGC CGCCAGCCGCUCACCUCCUUGGAG GCCAUGUGG |
| MIF1_DNA_primer | 5'-AGCGCAGACAGCGTGGGTCCCTG CGGCTCTTAGG |
| MIF1_RNA_primer | 5'-AGCGCAGACAGCGUGGGUCCCUG CGGCUCUUAGG |
| MIF1_RNA_template_blunt | 5'-GUGAGCGGCUGGCGGCUGUUCAA GAAGAUUAGCUAACCUAAGAGCCGC AGGGACCCACGCUGUCUGCGCU |
| MIF1_DNA_template_blunt | 5'-GTGAGCGGCTGGCGGCTGTTCAA GAAGATTAGCTAACCTAAGAGCCGC AGGGACCCACGCTGTCTGCGCT |
| MIF1_DNA_Y-shaped | 5'-GTCGTCCGCGGAGCACATGACGCTG TCTGCGCT |
| MIF1_Boltz2_DNA | 5'-GTCCCTGCGGCTCTTAGG |
| MIF1_Boltz2_RNA | 5'-AGCUAACCUAAGAGCCGCAGGGAC |
| Test_substrate_primer[103] | 5'-CGACGTCGCAG |
| Test_substrate_template[103] | 5'-UUUUUUUCGCGCUGCGACGUCG |

**Table A.1:** Oligo Sequences without labels.

# B

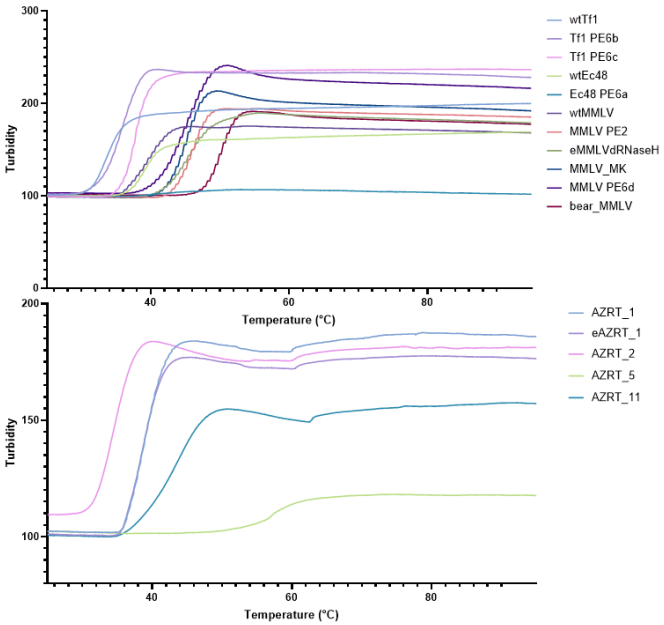
# Complementary Experiments

## B.1 Aggregation

Aggregation of the RTs were estimated using turbidity measurements which plots the average turbidity against the temperature (*figure B.1A*). The curves were averaged based on four biological replicates. The graphs show that the turbidity of Ec48 PE6a did not increase with temperature, either due to no aggregation tendencies (highly unlikely), or unknown issues with the sample affecting the aggregation. AZRT_5 has very high aggregation temperature, one possible explanation to this is that the concentration was too low in the sample to accurately track the turbidity.

Size of the particles were also measured simultaneously using DLS. The output was poly dispersity index (PDI) which is an estimation of how disperse the particles in the solution are (*figure B.1B*). PDI $< 0.1$: Monodisperse, $0.1 <$ PDI $< 0.4$: Moderately polydisperse and PDI $> 0.4$: Polydisperse. This is also an indication of aggregation and instability, as a polydisperse solution indicates particles of many different sizes - aggregation. The results showed that the MMLV RT family to a high extent are monodisperse while the Tf1 family are moderately polydisperse along with AZRT_1, eAZRT_1 and AZRT_11. AZRT_2 and AZRT_5 showed polydispersity at 25°C. The measurement of Ec48 PE6a is probably not representative of the RT as there was indication that the purification of the RT was unreliable.

**A.**




**B.**

| RT | avgPDI | Average radius (nm) |
|---|---|---|
| wtTf1 | 0,20 | 18,89 |
| Tf1 PE6c | 0,31 | 22,41 |
| Tf1 PE6c | 0,20 | 14,50 |
| wtEc48 | 0,19 | 17,89 |
| Ec48 PE6a | 0,40 | 2,91 |
| wtMMLV | 0,11 | 3,98 |
| MMLV PE2 | 0,08 | 4,03 |
| MMLVdRNaseH | 0,09 | 3,45 |
| MMLV_MK | 0,09 | 4,02 |
| MMLV PE6 | 0,07 | 3,40 |
| bear_MMLV | 0,11 | 5,70 |

| RT | avrg. radius | PDI |
|---|---|---|
| AZRT_1 | 6,4 | 0,23 |
| eAZRT_1 | 7,1 | 0,22 |
| AZRT_2 | 104,6 | 1,55 |
| AZRT_5 | 3,1 | 0,94 |
| AZRT_11 | 5,7 | 0,20 |

**Figure B.1:** **A.** Turbidity curves of purified RTs. **B.** PDI of purified RTs at 25°C.

## B.2 Thermostability

Thermostability was estimated using two different methods to obtain melting curves: tryptophan fluorescence (*figure B.2*) and thermal shift assay (*figure B.3*). The thermal shift assay showed that wtMMLV and MMLV PE2 might have two different melting events taking place, with two separate Tm. The same might apply to wtTf1 as well. AZRT_5 could not be assessed using this method due to low stock concentration.

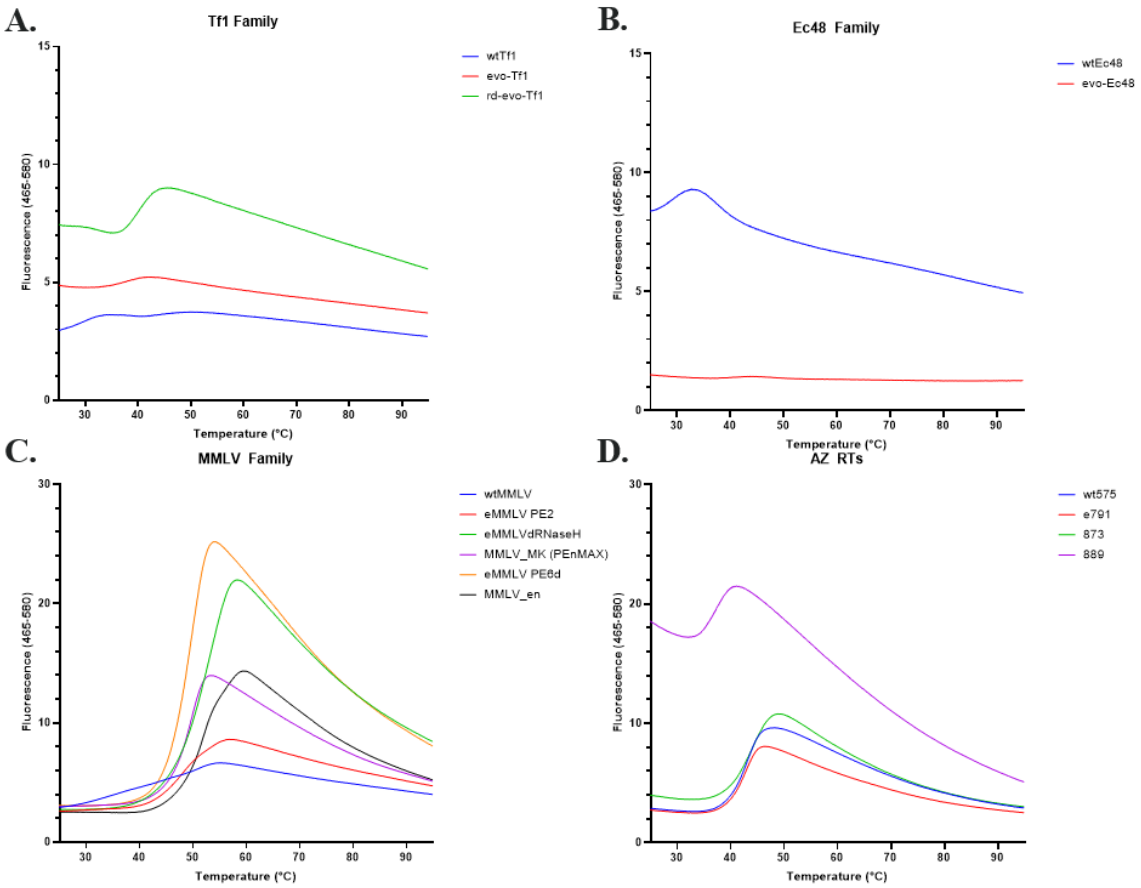



**Figure B.2:** Thermal shift melting curves of purified RTs. **A.** Tf1 RT Family. **B.** Ec48 RT Family. **C.** MMLV RT Family. **D.** AZ RTs, excluding AZRT_5.

Melting point was also estimated using tryptophan fluorescence. The instrument produced melting points with low variance between biological replicates (*figure B.3B*). However, the melting point of Ec48 PE6a could not be determined using this method due to noisy signal indicating issues with tryptophan fluorescence.

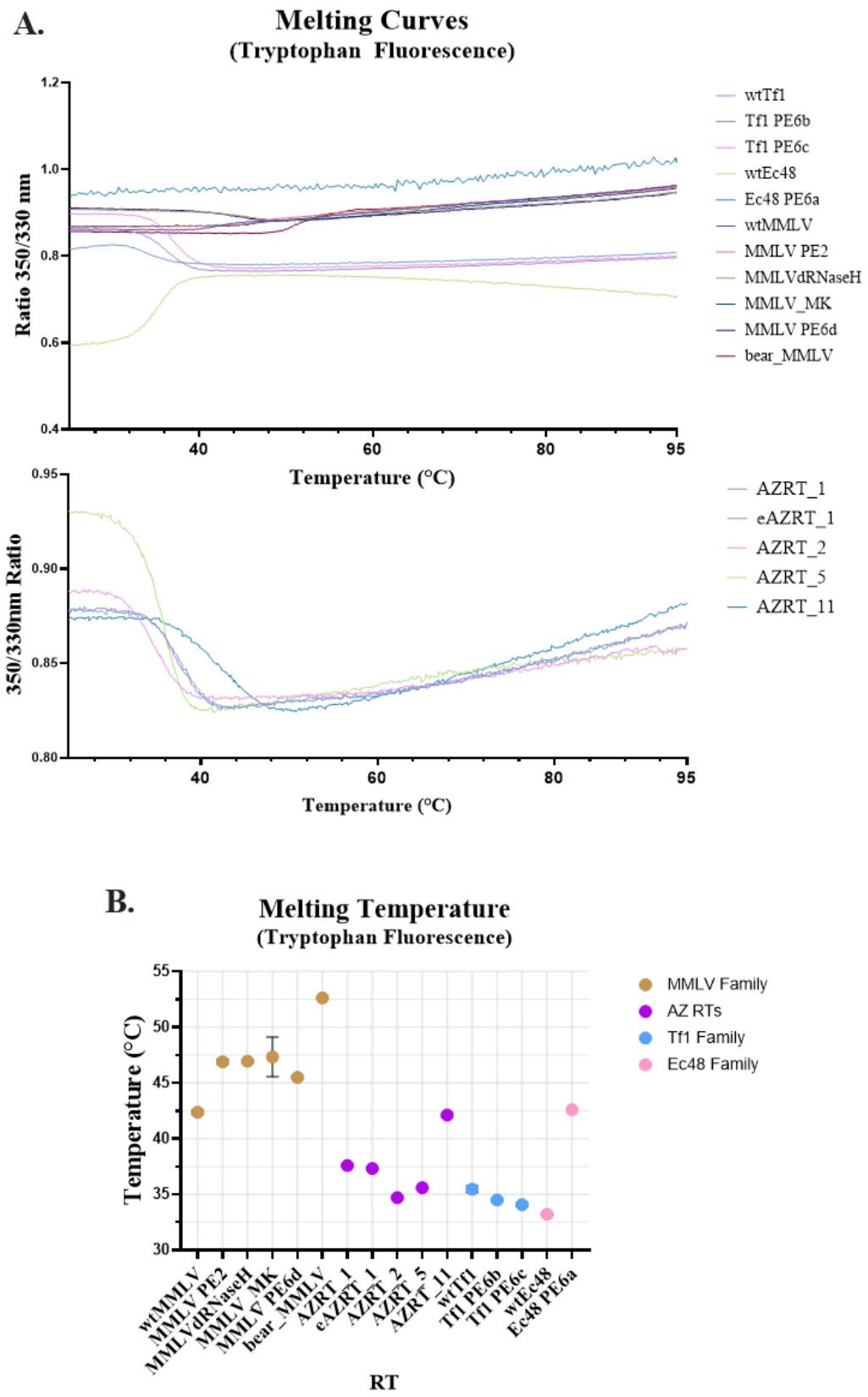



**Figure B.3:** **A.** Tryptophan fluorescence melting curves of purified RTs. **B.** Melting points of purified RTs from tryptophan fluorescence measurements.

The thermostability assays produced slightly different melting points for the RTs (*table B.1*). Some RTs seem to be more stable when measuring tryptophan fluorescence (wtMMLV, bear_MMLV, wtTf1, wtEc48, AZRT_11) which is probably an indication that the buffer conditions used for the tryptophan fluorescence measurements were more favorable for these RTs. The other RTs are more thermostable with the thermal shift assay. The largest difference is observed with the AZ RTs which show a difference between 5 and 8.6° between methods in favor of thermal shift, except AZRT_11. wtTf1 also show a significant difference with Tm being 35.5°C with tryptophan fluorescence, and 28.8°C with thermal shift. In general, patterns withing RT families seem to be the same with wtRTs having lower Tm than the engineered counterparts. Comparisons between families also produce similar correlations. For example, the AZRTs seem to be similarly stable to the engineered Tf1 RTs, and the MMLV RTs are the most thermostable regardless of method of measurement.

| **RT** | **Tm Tryptophan Fluorescence [°C]** | **Tm Thermal Shift [°C]** |
|---|---|---|
| wtMMLV | 42.4 | 41.1 |
| MMLV PE2 | 46.9 | 48.1 |
| MMLVΔRNaseH | 46.9 | 52.1 |
| MMLV_MK | 47.3 | 49.2 |
| MMLV PE6d | 45.5 | 49.2 |
| bear_MMLV | 52.6 | 51.8 |
| wtTf1 | 35.5 | 28.8 |
| Tf1 PE6b | 34.5 | 39.5 |
| Tf1 PE6c | 34.5 | 40.3 |
| wtEc48 | 33.2 | 28.7 |
| Ec48 PE6a | 42.6* | 40.5 |
| AZRT_1 | 37.6 | 42.6 |
| eAZRT_1 | 37.3 | 42.3 |
| AZRT_2 | 34.7 | 43.3 |
| AZRT_5 | 35.6 | - |
| AZRT_11 | 42.1 | 37.0 |

**Table B.1:** Comparison of mean melting points measured using thermal shift assay and tryptophan fluorescence. Means are calculated from n=4 biological replicates. *Unreliable

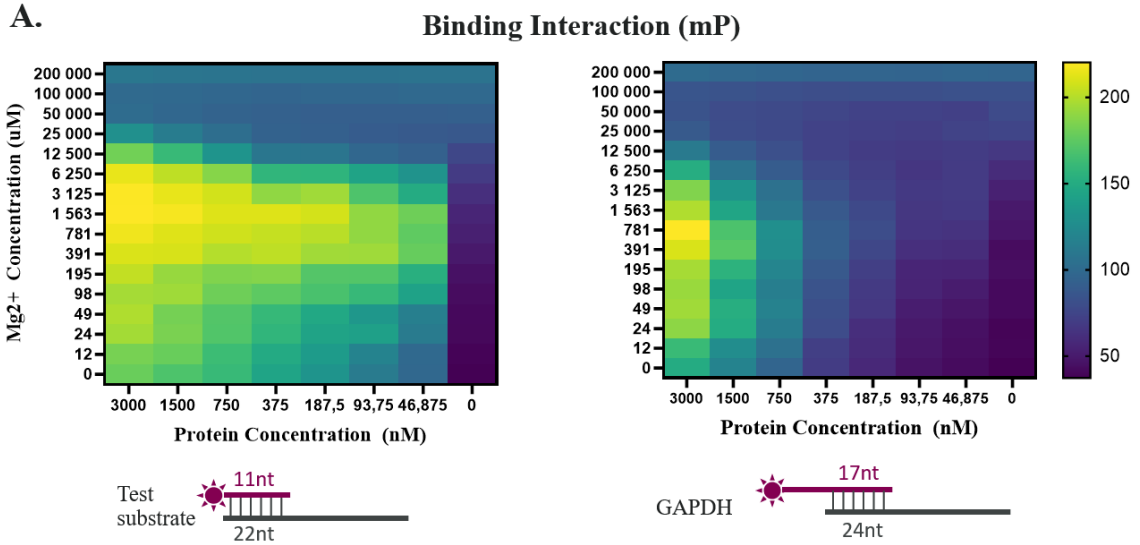
## B.3 Binding

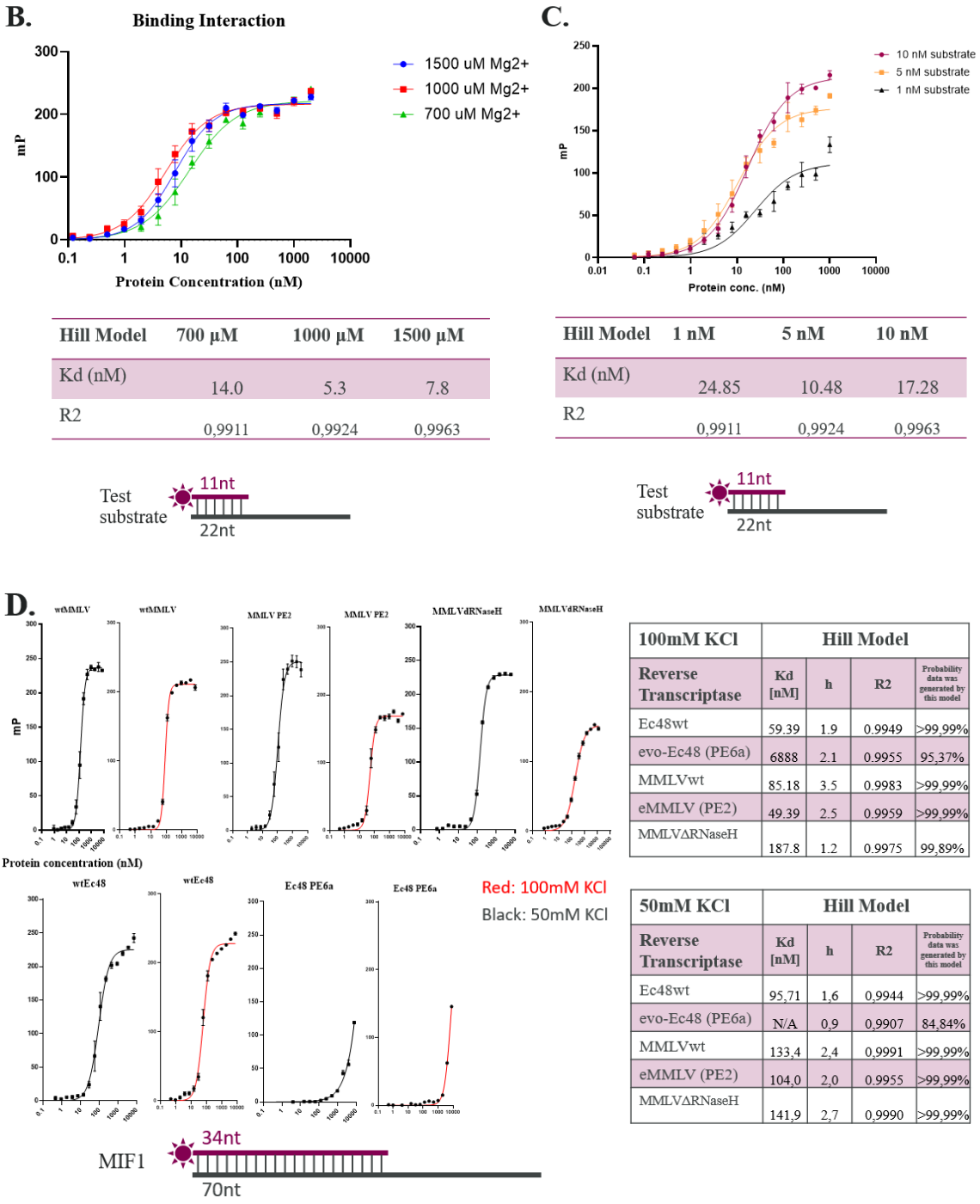
To optimize the binding assay multiple conditions were tested. The $MgCl_2$ concentration was first optimized by titrating $MgCl_2$ against protein concentration in

the fluorescent polarization assay using two different substrates (*figure B.4A*). The result showed that binding interaction seemed to be stronger in the range around 1 μM $MgCl_2$, indicated by higher mP values. The data also revealed that too high $MgCl_2$ concentration seemed to cause the light to stay polarized even without protein. Possibly it caused the substrate to form aggregates. To decide a specific $MgCl_2$ to proceed with for the assay, three different concentrations of $MgCl_2$ was selected, 1.5 mM, 1 mM, and 0.75 mM (*figure B.4B*). The result showed that within the range was optimal to promote binding interaction, and 1 mM was thus selected for assays.

The optimization of $MgCl_2$ was done with 10 nM substrate concentrations, however, it was speculated that a lower concentration of the substrate could be used to obtain Kd on the single nM scale. Therefore, three different concentrations of substrate were tested in the fluorescence polarization assay using 1 mM $MgCl_2$ (*figure B.4C*). The hill constant was locked to 1, meaning noncooperative binding. The results showed that while 1 nM resulted in slightly more variation and in total lower levels of light polarization, the instrument was sensitive enough to detect the polarization of the light and the data fit well to the hill model.

Next, two different buffer conditions were tested with the fluorescent polarization (*figure B.4D*). It was hypothesized that the high salt concentration in the buffer might reduce binding interaction and that a lower salt concentration might increase sensitivity, and allow complete binding curves for RTs that did not reach the plateau due to the protein stock limiting the maximum protein concentration that could be used for the assay. The result showed that lowering the salt concentration allowed a higher maximum binding, however, it increased Kd. Most importantly, it did not allow for completion of the binding curves with the RTs that did not reach a plateau. It was decided to continue using 100 mM KCl in the reaction buffer.






| Hill Model | 700 µM | 1000 µM | 1500 µM |
|---|---|---|---|
| Kd (nM) | 14.0 | 5.3 | 7.8 |
| R2 | 0,9911 | 0,9924 | 0,9963 |

| Hill Model | 1 nM | 5 nM | 10 nM |
|---|---|---|---|
| Kd (nM) | 24.85 | 10.48 | 17.28 |
| R2 | 0,9911 | 0,9924 | 0,9963 |

| 100mM KCl | Hill Model | | | |
|---|---|---|---|---|
| Reverse Transcriptase | Kd [nM] | h | R2 | Probability data was generated by this model |
| Ec48wt | 59.39 | 1.9 | 0.9949 | >99,99% |
| evo-Ec48 (PE6a) | 6888 | 2.1 | 0.9955 | 95,37% |
| MMLVwt | 85.18 | 3.5 | 0.9983 | >99,99% |
| eMMLV (PE2) | 49.39 | 2.5 | 0.9959 | >99,99% |
| MMLVΔRNaseH | 187.8 | 1.2 | 0.9975 | 99,89% |

| 50mM KCl | Hill Model | | | |
|---|---|---|---|---|
| Reverse Transcriptase | Kd [nM] | h | R2 | Probability data was generated by this model |
| Ec48wt | 95,71 | 1,6 | 0,9944 | >99,99% |
| evo-Ec48 (PE6a) | N/A | 0,9 | 0,9907 | 84,84% |
| MMLVwt | 133,4 | 2,4 | 0,9991 | >99,99% |
| eMMLV (PE2) | 104,0 | 2,0 | 0,9955 | >99,99% |
| MMLVΔRNaseH | 141,9 | 2,7 | 0,9990 | >99,99% |

**Figure B.4:** **A.** $MgCl_2$ titration agains protein concentration during fluorescence polarization. **B.** Three $MgCl_2$ tested during fluorescent polarization and fitted to the hill model. **C.** Three different substrate concentrations tested during fluorescent polarization and fitted to the hill model with the hill constant locked to 1. **D.** Two different buffer conditions tested with the fluorescent polarization assay, 50 mM KCl and 100 mM KCl.

After confirming the conditions for the polymerization assay. The RTs were benchmarked using the MIF1 substrate. The data was fitted to two different binding

interaction models to determine which one described the binding interaction the best (*table B.2*). The hill constant was not locked, to allow the model to estimate the constant. The results showed that most RTs fit the Hill model the best, with GraphPad Prism estimating that the probability of the Hill Model generating the data was close to 100%. Exceptions were bear_MMLV, eAZRT_1, AZRT_5. The result also showed that many of the RTs seem to interact with positive cooperation to the substrate, indicated by hill constant h>1.

| MIF1 | Hill Model | | | | Two Sites Model | | | |
|---|---|---|---|---|---|---|---|---|
| Reverse Transcriptase | Kd [nM] | h | R2 | Probability data was generated by this model | KdHi [nM] | KdLo [nM] | R2 | Probability data was generated by this model |
| Tf1wt | 101.6 | 2.2 | 0.9979 | >99,99% | 128.2 | 128.2 | 0.9723 | <0,01% |
| Tf1 PE6b | 91.82 | 2.5 | 0.9944 | >99,99% | 120.5 | 120.5 | 0.9662 | <0,01% |
| Tf1 PE6c | 51.39 | 1.9 | 0.9979 | >99,99% | 62.32 | 62.32 | 0.9786 | <0,01% |
| Ec48wt | 59.39 | 1.9 | 0.9949 | >99,99% | 71.79 | 71.82 | 0.9787 | <0,01% |
| Ec48 PE6a | 6888 | 2.1 | 0.9955 | 95,37% | 10994 | 12829 | 0.9950 | 4,626% |
| MMLVwt | 85.18 | 3.5 | 0.9983 | >99,99% | N/A | 103.3 | 0.9477 | <0,01% |
| MMLV PE2 | 49.39 | 2.5 | 0.9959 | >99,99% | 56.57 | 56.57 | 0.9636 | <0,01% |
| MMLVΔRNaseH | 187.8 | 1.2 | 0.9975 | 99,89% | 204.1 | 204.1 | 0.9955 | 0,1112% |
| MMLV_MK | 61.28 | 2.6 | 0.9939 | >99,99% | N/A | 65.54 | 0.9527 | <0,01% |
| MMLV PE6d | 327.5 | 1.2 | 0.9976 | 99,92% | 369.4 | 369.4 | 0.9955 | 0,08172% |
| bear_MMLV | 111.2 | 1.0 | 0.9814 | 18,4% | 83.34 | 10656 | 0.9882 | 81,6% |
| AZRT_1 | 1929 | 1.0 | 0.9989 | 93,99% | 2046 | 2046 | 0.9988 | 6,008% |
| eAZRT_1 | 3492 | 0.8 | 0.9986 | 0,0323% | 2.338 | 2770 | 0.9996 | 99,97% |
| AZRT_2 | 4135 | 0.8 | 0.9978 | 79,64% | 40.09 | 3383 | 0.9980 | 20,36% |
| AZRT_5 | 347.8 | 0.7 | 0.9826 | 0,4763% | 0.2294 | 258.4 | 0.9932 | 99,52% |
| AZRT_11 | 750.5 | 1.2 | 0.9979 | 99,97% | 908.3 | 908.3 | 0.9956 | 0,02954% |

**Table B.2:** Comparison of the Hill model and Two-sites binding model on fluorescent polarization data obtained from binding interactions with the MIF1 substrate.

The same benchmarking was doen for the GAPDH substrate (*figure B.3*). Again, Hill model provided the best fit of the data, with bear_MMLV being an exception once again. AZRT_2 also fit the two site model the best.

| GAPDH | Hill Model | | | | Two Sites Model | | | |
|---|---|---|---|---|---|---|---|---|
| Reverse Transcriptase | Kd [nM] | h | R2 | Probability data was generated by this model | KdHi [nM] | KdLo [nM] | R2 | Probability data was generated by this model |
| Tf1wt | 282,2 | 3,0 | 0,9984 | >99,99% | 407,7 | 407,7 | 0,9566 | <0,01% |
| Tf1 PE6b | 227,3 | 3,5 | 0,9946 | >99,99% | 308,2 | 308,2 | 0,9534 | <0,01% |
| Tf1 PE6c | 118,2 | 4,9 | 0,9965 | >99,99% | 166,4 | 166,4 | 0,9387 | <0,01% |
| Ec48wt | 79,99 | 2,6 | 0,9987 | >99,99% | 102,6 | 102,7 | 0,9606 | <0,01% |
| Ec48 PE6a | 492501 | 2,0 | 0,9855 | 81,25% | 11591494 | 51513038 | 0,9867 | 18,75% |
| MMLVwt | 332,5 | 1,4 | 0,9931 | | | | | |
| MMLV PE2 | 254,1 | 1,9 | 0,9977 | >99,99% | 358,0 | 358,0 | 0,9764 | <0,01% |
| MMLVΔRNaseH | 305,0 | 1,9 | 0,9977 | >99,99% | 429,5 | 429,6 | 0,9764 | <0,01% |
| MMLV_MK | 318,1 | 1,7 | 0,9955 | >99,99% | 469,6 | 470,3 | 0,9795 | <0,01% |
| MMLV PE6d | 540,5 | 1,4 | 0,9974 | >99,99% | 663,8 | 664,7 | 0,9904 | <0,01% |
| bear_MMLV | 1602 | 0,8 | 0,9939 | 3,351% | 707,4 | N/A | 0,9969 | 96,65% |
| AZRT_1 | 3275 | 1,0 | 0,9973 | 90,13% | 3184 | 3184 | 0,9973 | 9,874% |
| eAZRT_1 | 3617 | 1,5 | 0,9975 | >99,99% | 6861 | 6862 | 0,9897 | <0,01% |
| AZRT_2 | 2276 | 0,9 | 0,9936 | 45,06% | 0,1863 | 2133 | 0,9953 | 54,94% |
| AZRT_5 | 288,5 | 1,3 | 0,9620 | 95,71% | N/A | ~ 363,5 | 0,9574 | 4,289% |
| AZRT_11 | 5053 | 0,8 | 0,9980 | 76,77% | 733,4 | 11510 | 0,9982 | 23,23% |

**Table B.3:** Comparison of the Hill model and Two-sites binding model on fluorescent polarization data obtained from binding interactions with the GAPDH substrate.

## B.4 Reverse transcription activity

Polymerization activity was measured through polymerization reactions of 20 seconds. The products of the repeats were pooled and visualized on a UREA gel (*figure B.5*).

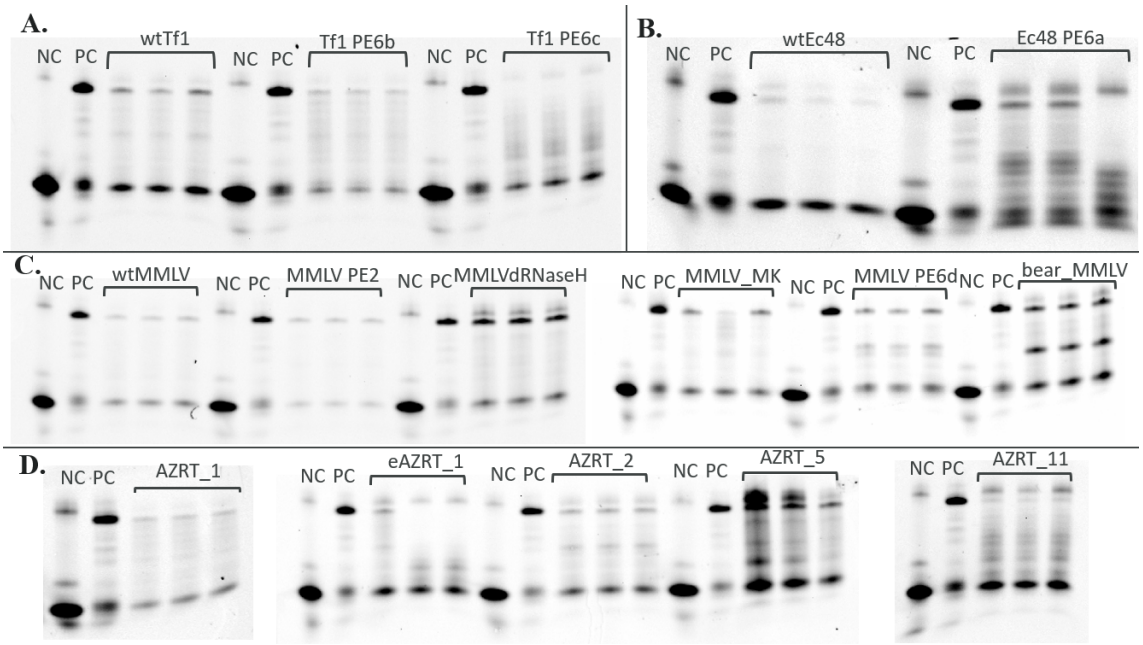



**Figure B.5: A.** Benchmarking polymerization activity of Tf1 RT family during 20 second reactions with three replicates. **B.** Benchmarking polymerization activity of Ec48 RT family during 20 second reactions with three replicates. **C.** Benchmarking polymerization activity of MMLV RT family during 20 second reactions with three replicates. **D.** Benchmarking polymerization activity of AZ RTs during 20 second reactions with three replicates.

The following figures show how the area of extended products was defined for each replicate, and what the program determined the area of each field as (*figure B.6-B.12*).

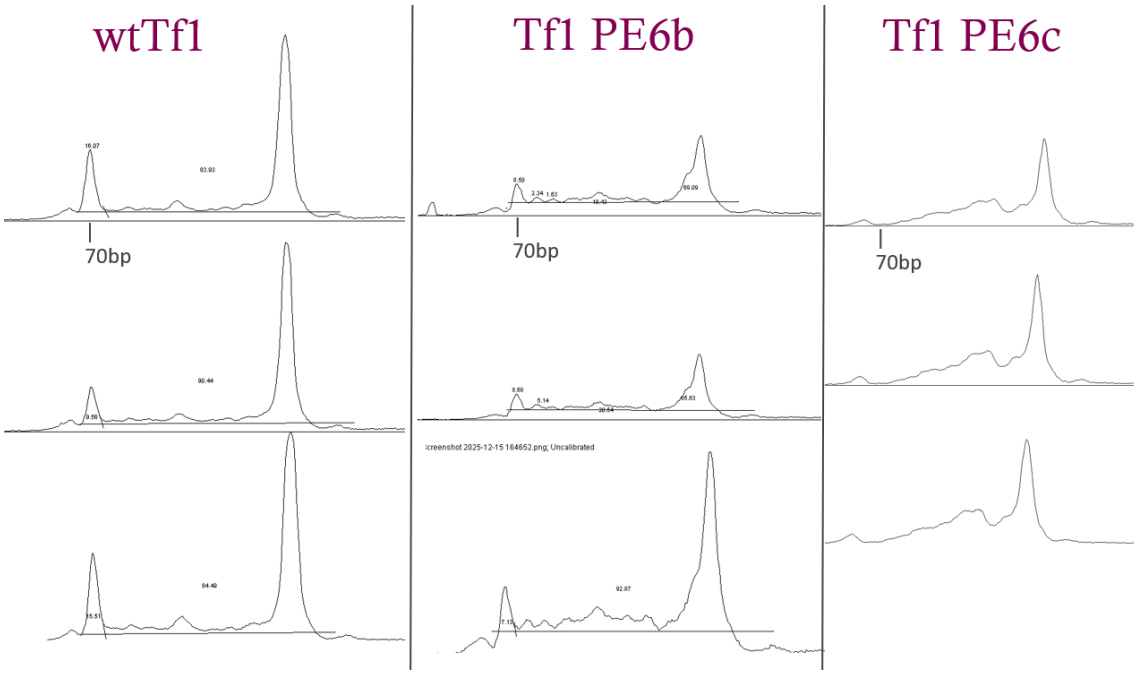



**Figure B.6:** Band intensity of wtTf1, Tf1 PE6b and Tf1 PE6c polymerization products from polymerization assay, processed in ImageJ.

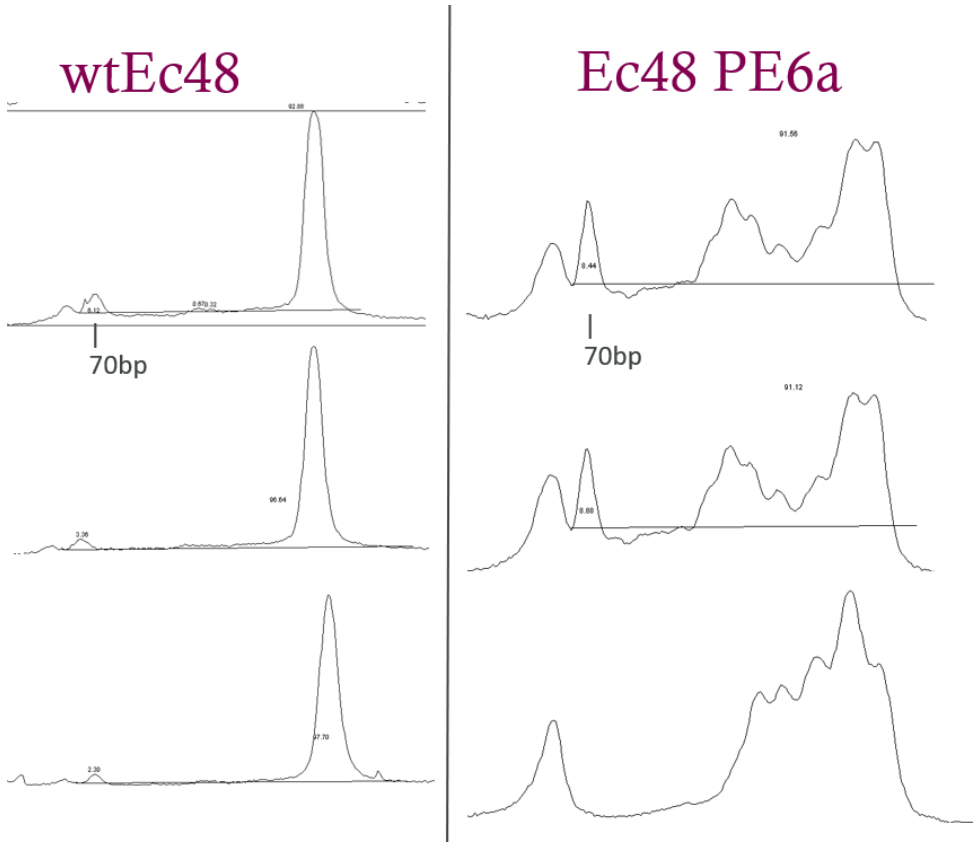



**Figure B.7:** Band intensity of wtEc48 and Ec48 PE6a polymerization products from polymerization assay, processed in ImageJ.

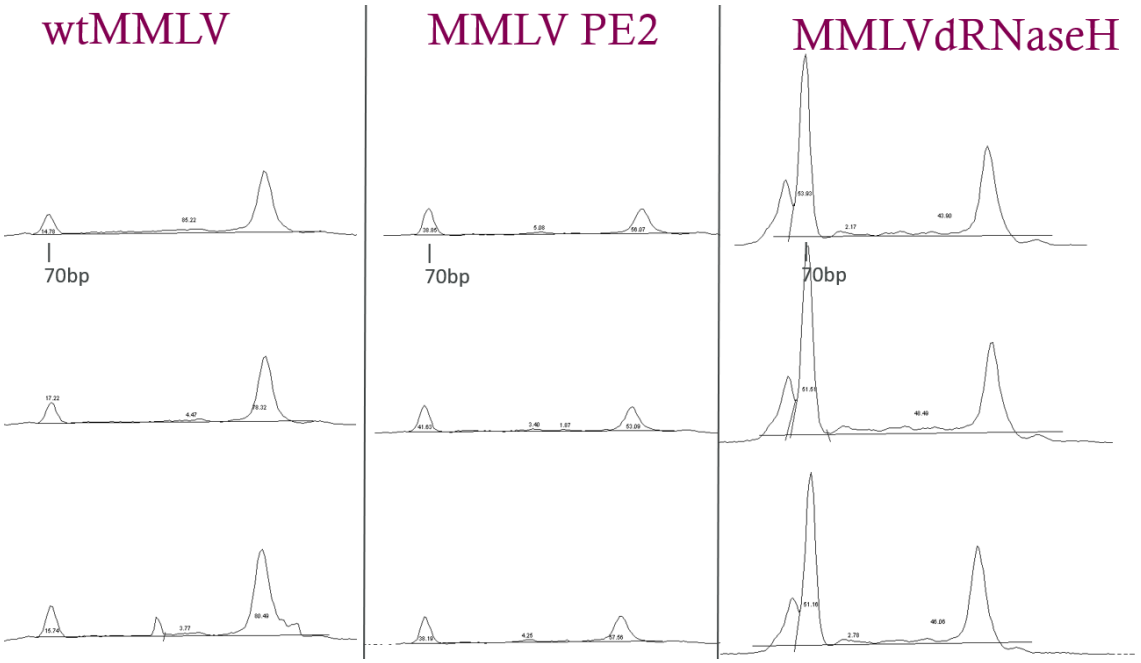



**Figure B.8:** Band intensity of wtMMLV, MMLV PE2 and MMLVΔRNaseH polymerization products from polymerization assay, processed in ImageJ.

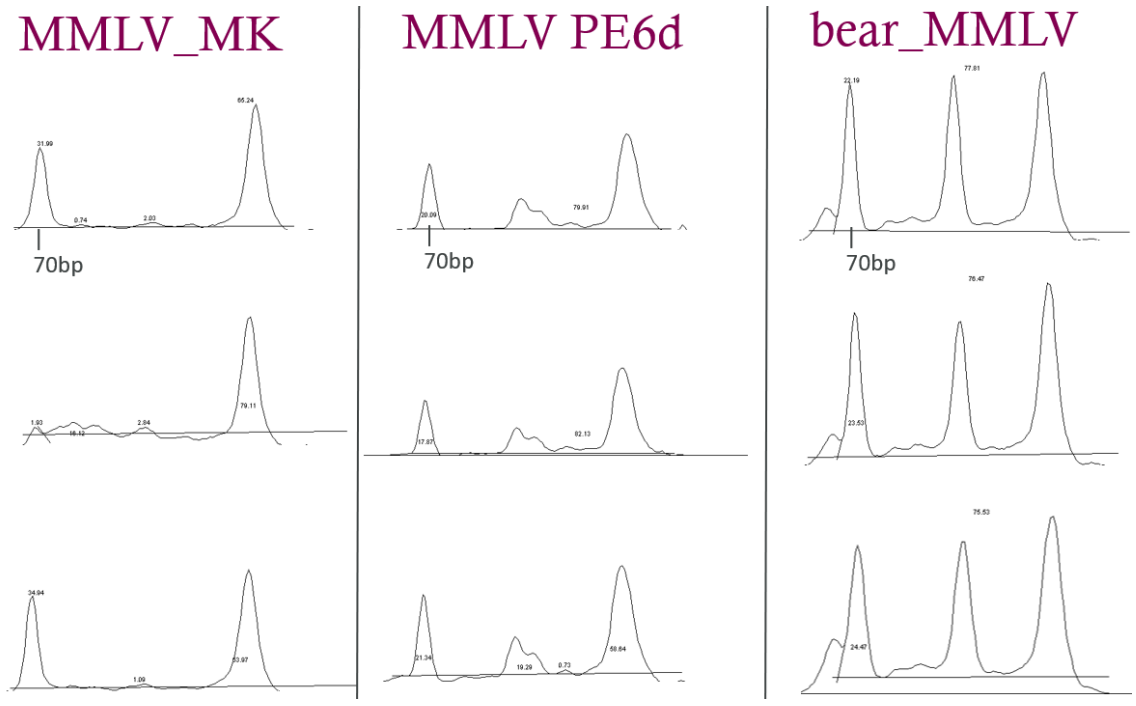



**Figure B.9:** Band intensity of MMLV_MK, MMLV PE6d and bear_MMLV polymerization products from polymerization assay, processed in ImageJ.

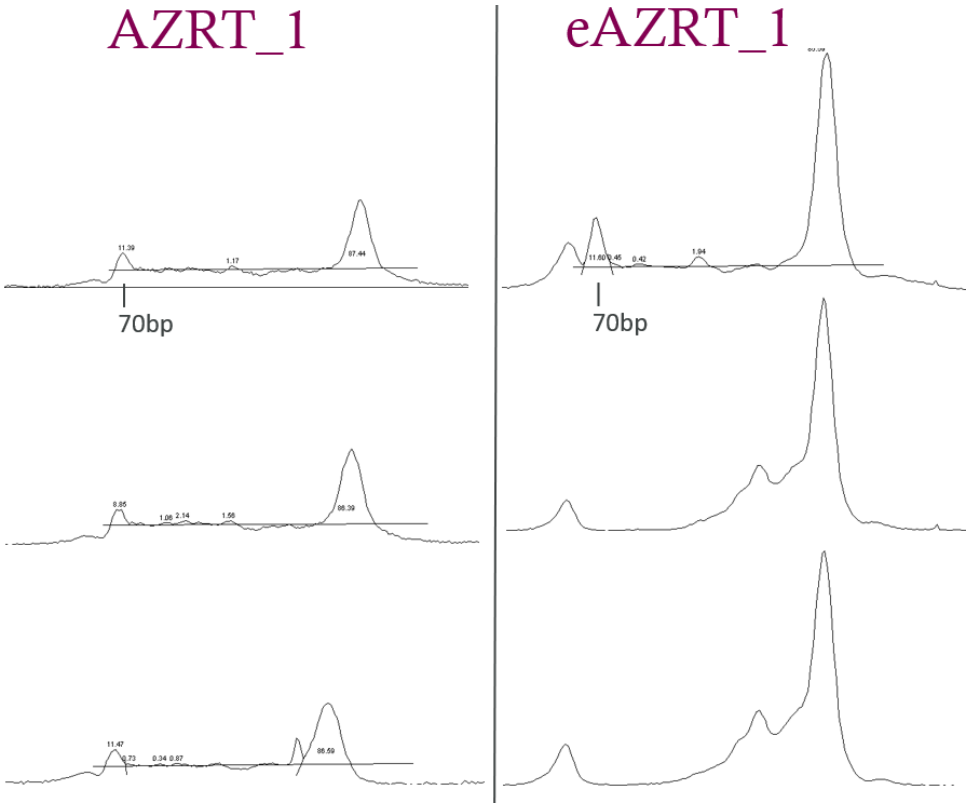



**Figure B.10:** Band intensity of AZRT_1 and eAZRT_1 polymerization products from polymerization assay, processed in ImageJ.

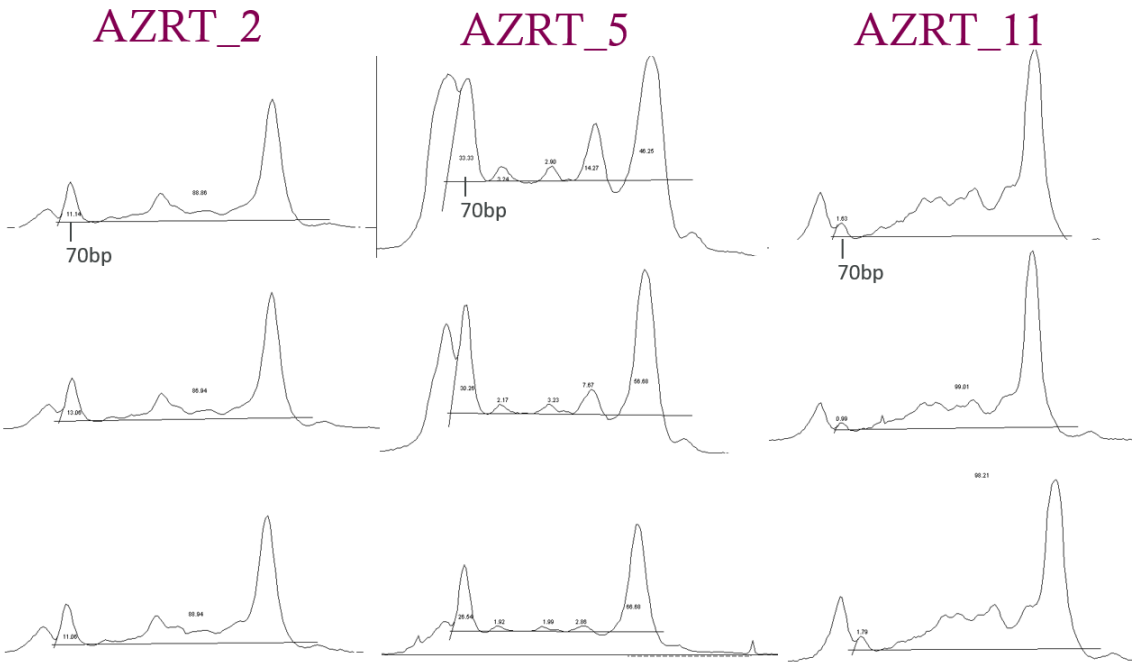



**Figure B.11:** Band intensity of AZRT_2, AZRT_5 and AZRT_11 polymerization products from polymerization, assay processed in ImageJ.

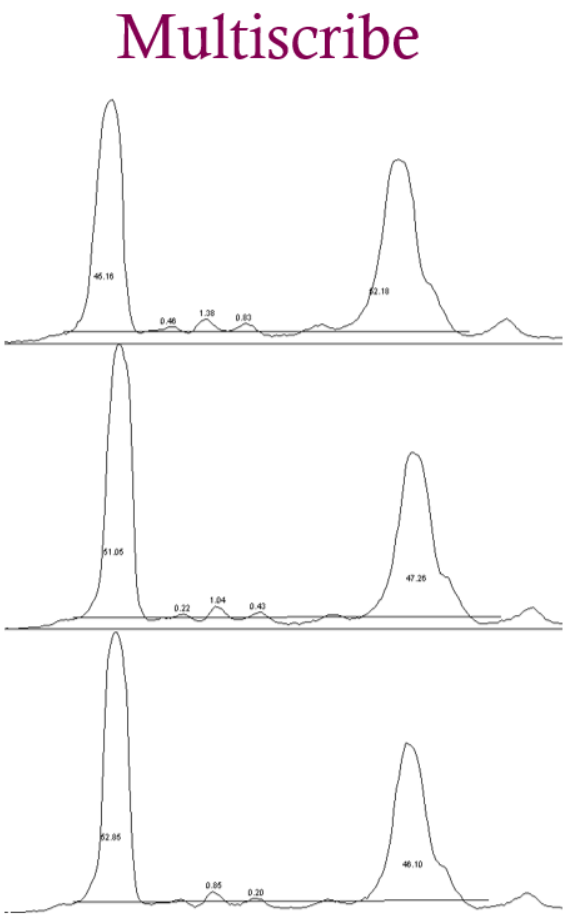



**Figure B.12:** Representative example of band intensity of Multiscribe RT polymerization products from polymerization assay, processed in ImageJ.

Raw data collected from the areas are collected in *table B.4*. Each RT was normalized using the replicates of the positive control which defined what was counted as a "completed reaction".

| **RT** | **rep. 1** [%] | **rep. 2** [%] | **rep. 3** [%] | **Mean** [%] | **Normalized** [%] |
|---|---|---|---|---|---|
| wtTf1 | 16,07 | 9,56 | 15,51 | 13,71 | 27,60 |
| Tf1 PE6b | 8,5 | 8,68 | 7,13 | 8,10 | 16,31 |
| Tf1 PE6c | 0 | 0 | 0 | 0 | 0 |
| PC | 45,16 | 51,05 | 52,85 | 49,69 | 1 |
| wtMMLV | 14,776 | 17,219 | 15,742 | 15,91 | 26,40 |
| MMLV PE2 | 38,851 | 41,632 | 38,185 | 39,56 | 65,62 |
| MMLVΔRNaseH | 53,927 | 51,513 | 51,161 | 52,20 | 86,60 |
| PC | 57,672 | 60,416 | 62,75 | 60,28 | 1 |
| MMLV_MK | 31,99 | 1,93 | 34,935 | 22,95 | 40,00 |
| MMLV PE6d | 20,092 | 17,87 | 21,339 | 19,77 | 34,45 |
| bear_MMLV | 22,194 | 23,525 | 24,47 | 23,40 | 40,78 |
| PC | 55,07 | 57,637 | 59,423 | 57,38 | 1 |
| wtEc48 | 6,12 | 3,36 | 2,303 | 3,93 | 8,50 |
| Ec48 PE6a | 8,444 | 8,88 | 0 | 5,77 | 12,49 |
| AZRT_1 | 11,39 | 8,848 | 11,465 | 10,57 | 22,86 |
| PC | 43,807 | 46,633 | 48,261 | 46,23 | 1 |
| eAZRT_1 | 11,603 | 0 | 0 | 3,87 | 6,41 |
| AZRT_2 | 11,144 | 13,064 | 11,064 | 11,76 | 19,49 |
| AZRT_5 | 33,334 | 30,255 | 26,544 | 30,04 | 49,81 |
| PC | 55,36 | 62,748 | 62,86 | 60,32 | 1 |
| AZRT_11 | 1,627 | 0,987 | 1,791 | 1,47 | 2,43 |
| PC | 43,46 | - | - | 43,46 | 1 |

**Table B.4:** Percentage of the initial substrate extended after 20 seconds. Separated into RTs that were run on the same gel. PC=Positive Control. Calculated through band intensity in ImageJ.

Substrate degradation was tested by incubating the RTs with the substrate without any dNTPs for 3h (*figure B.13*). If there were any additional bands that appeared from this treatment not present in the NC, it suggested that the RT was degrading the substrate through nuclease activity, either from contamination of nuclease or intrinsic nuclease activity.

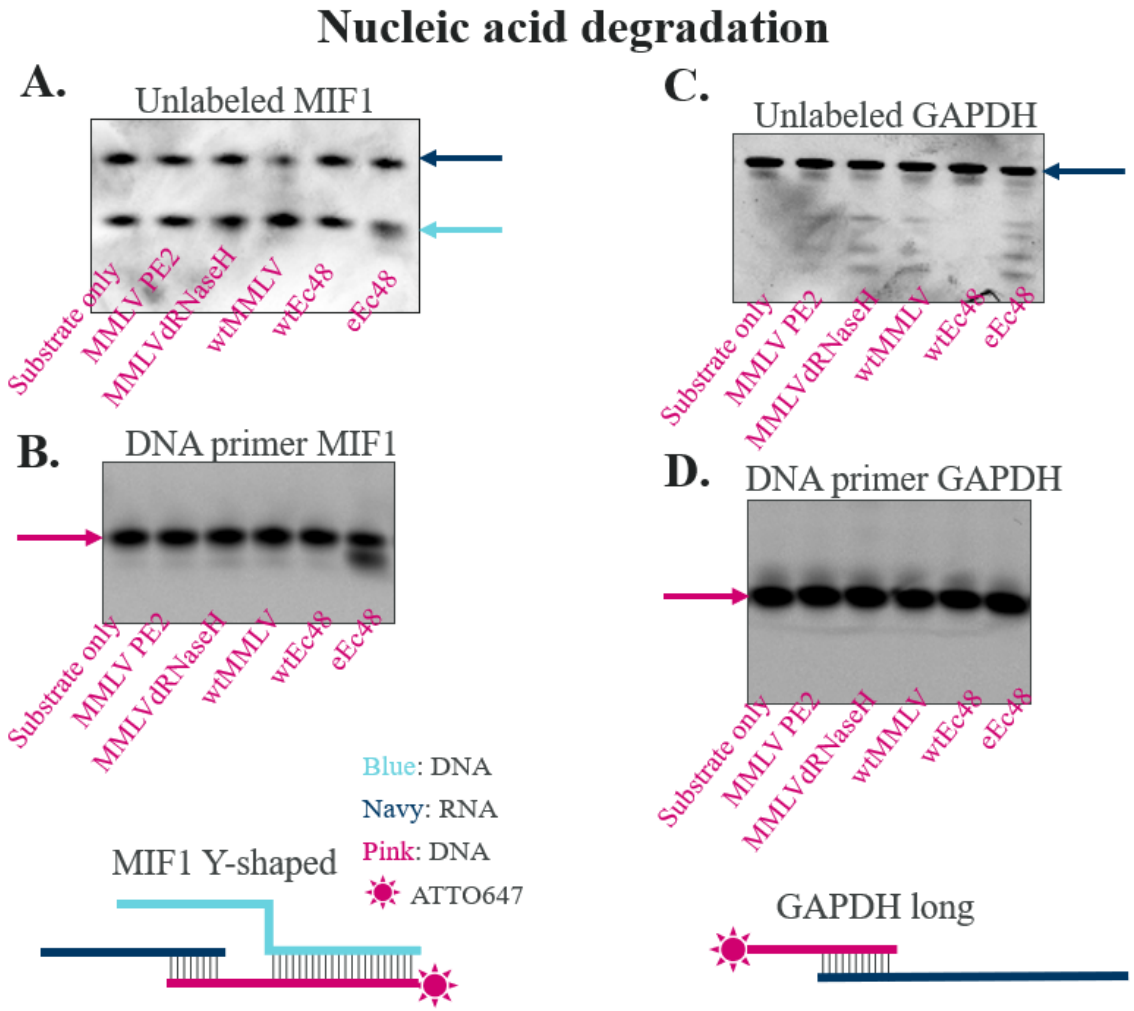



**Figure B.13:** **A.** Incubation of RTs together with unlabeled RNA and DNA of the MIF1 Y-substrate for 3 hours. **B.** Incubation of RTs together with labeled DNA primer of the MIF1 Y-substrate for 3 hours. **C.** Incubation of RTs together with unlabeled RNA and DNA of the GAPDH substrate for 3 hours. **D.** Incubation of RTs together with labeled DNA primer of the GAPDH substrate for 3 hours.

**DEPARTMENT OF SOME SUBJECT OR TECHNOLOGY**
**CHALMERS UNIVERSITY OF TECHNOLOGY**
Gothenburg, Sweden
www.chalmers.se


CHALMERS
UNIVERSITY OF TECHNOLOGY