Figure S1

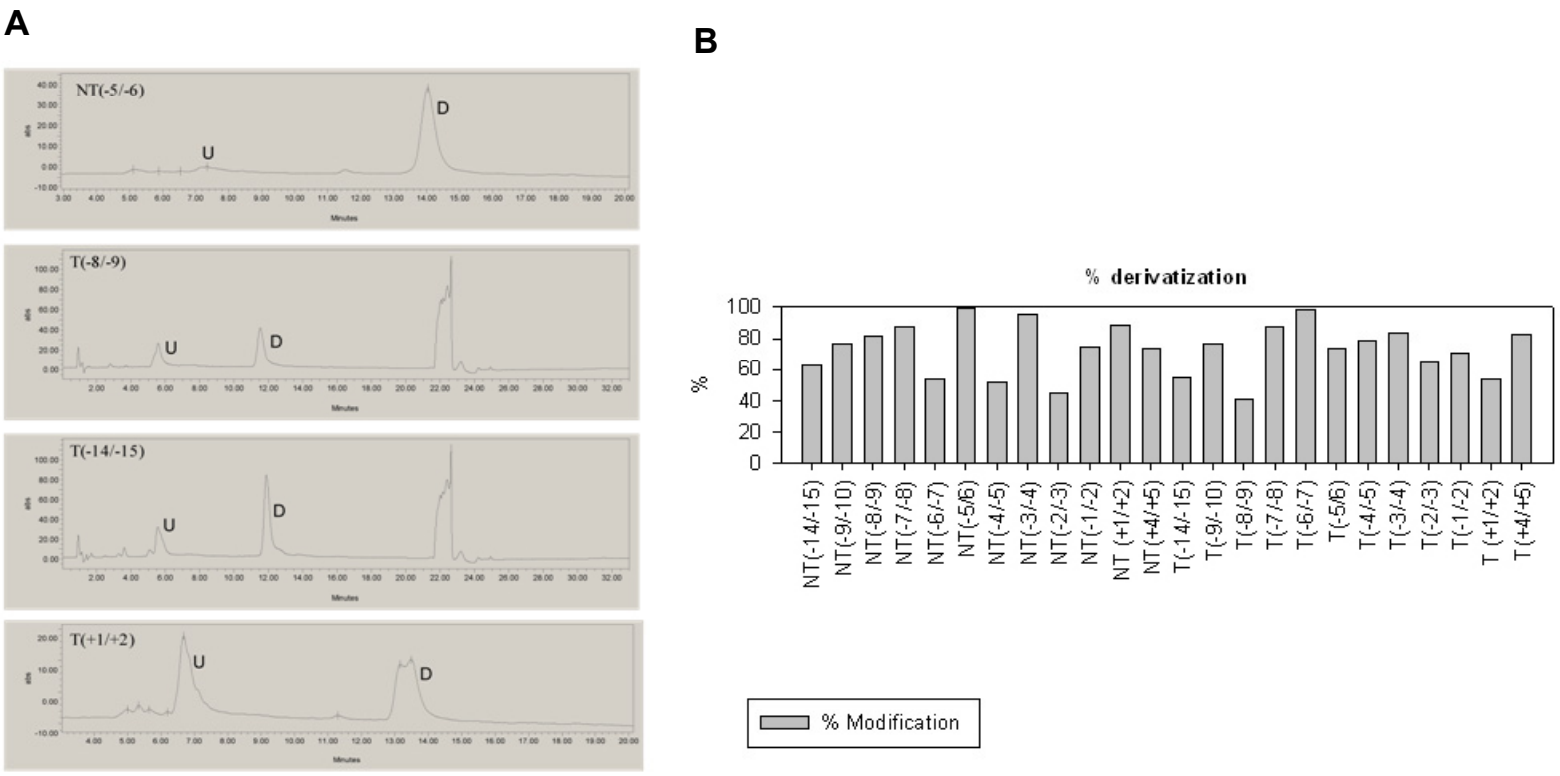


Figure S1: Derivatization of promoter DNAs

A) HPLC traces for some promoter DNAs used in the study: Promoter DNAs phosphorothioated at indicated positions were derivatized with 4-azidophenacyl bromide as described in Methods. Samples were analyzed using HPLC (C18 column). Peaks corresponding to underivatized DNAs, labeled “U” (Retention time around 6) and those corresponding to derivatized DNAs, labeled “D” (retention times around 12-15) were integrated using standard methods. % derivatization was calculated according to the following equation:

% derivatization= {Area of Derivatized DNA/(Area of underivatized DNA + derivatized DNA)}*100]

B): % derivatization for all the DNAs used in the study

Figure S2

A

```
NT   5′ CCATAATTTATTTATTATTATATAAGTAATAAATAATTGTTTTATATCC
 T       TATTAAATAAATAATAATATATTCATTATTTATTAACAAAATATAGGCC  5′
```

B

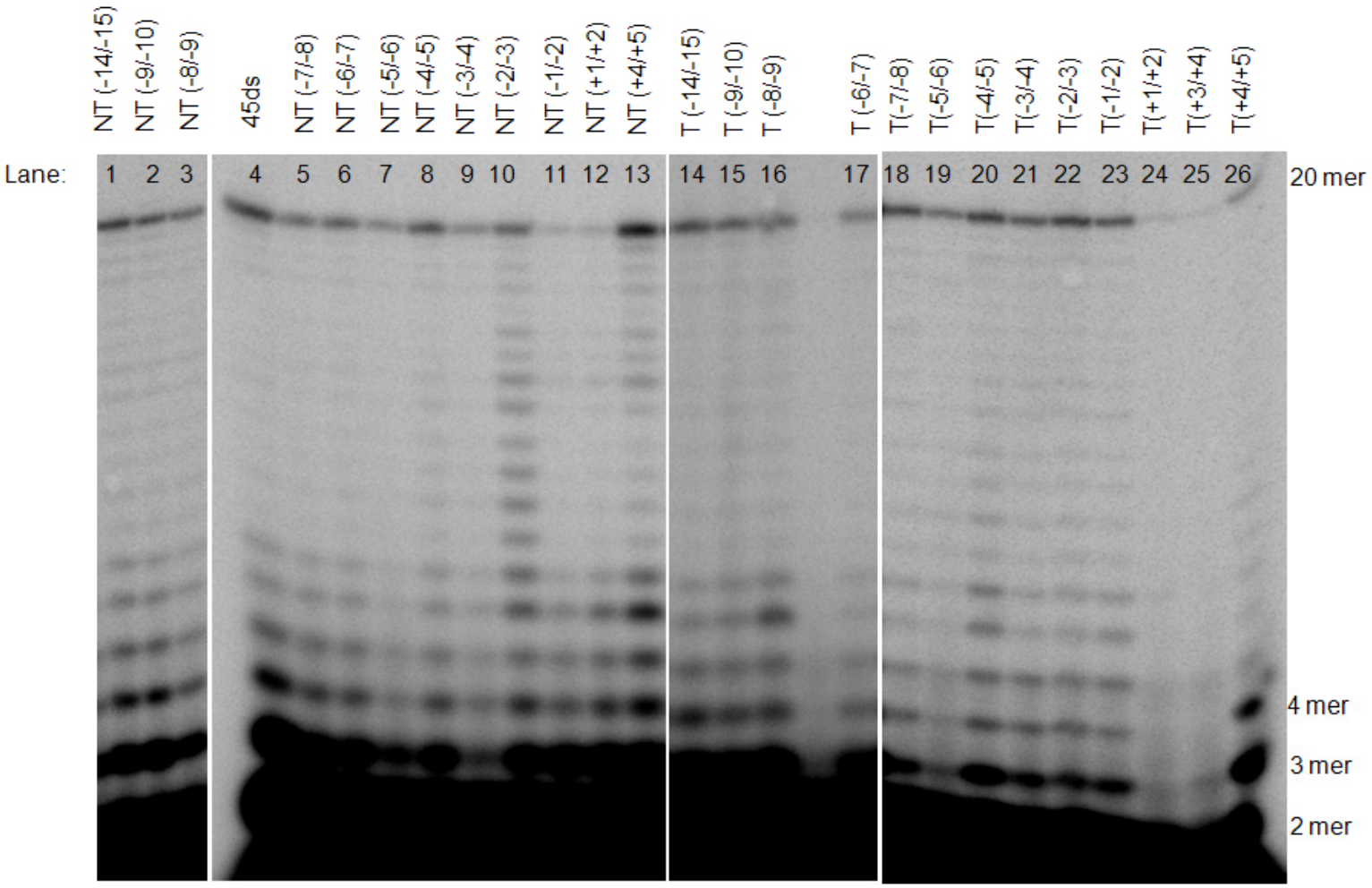


Figure S2: Transcription reaction products with the derivatized promoter DNAs
A: Duplex promoter sequence used for transcription studies. NT (non-template strand), T (template strand) B: Transcription reactions were carried outused for transcription reactions for 3 min at 22^{0}C with an equimolar mixture of Rpo41 and Mtf1 (500nM each), derivatized DNA (1μM), ATP, UTP and GTP (250μM each) spiked with γ[^{32}P]ATP. The RNA products were resolved on an 18% polyacrylamide sequencing gel containing 7 M urea. The first C in the DNA sequence is encountered at +21 (Fig. 2). Exclusion of CTP, therefore, results in a runoff product of 20 nt.

Figure S3

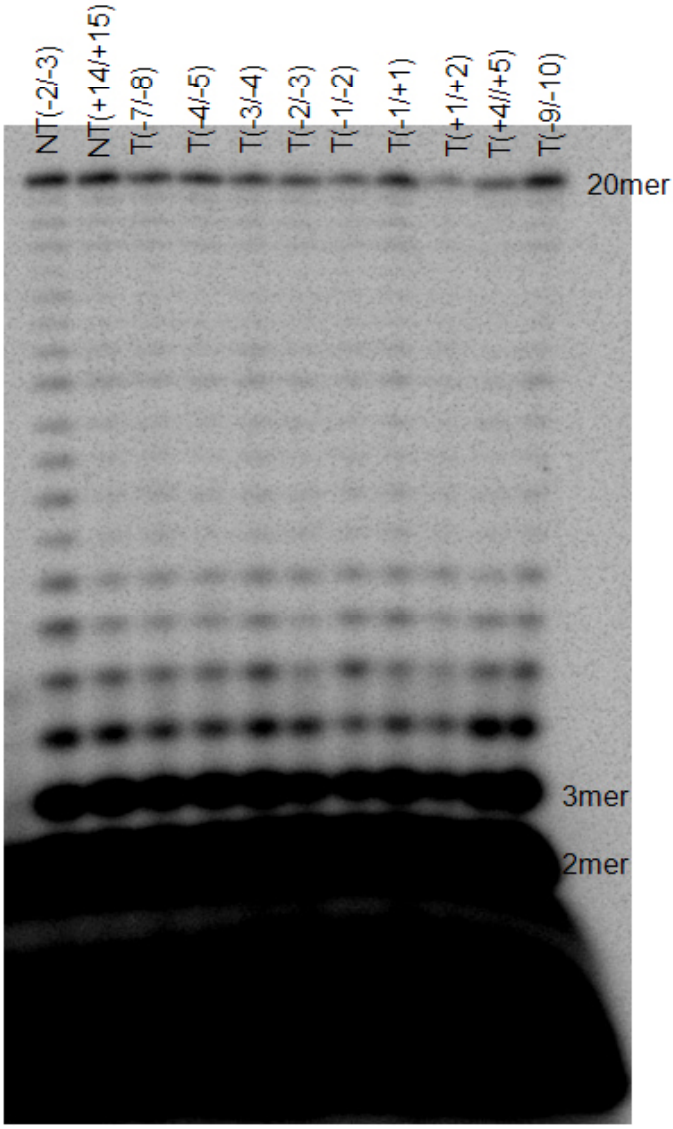


Figure S3: Transcription reaction with phosphorothioated DNAs
Equimolar mixture of Rpo41 and Mtf1 (500 nM each) was mixed with phosphorothioated DNA (1uM), ATP, UTP and GTP (250uM each) spiked with γ[^{32}P]ATP for 3 min at 22^{0}C. The RNA products were resolved on an 18% polyacrylamide sequencing gel containing 7 M urea. The gel was exposed to a storage phosphor screen and phosphorimaged. The first C in the DNA sequence is encountered at +21 (Fig. 2). Exclusion of CTP, therefore, results in a runoff product of 20 nt.

Figure S4

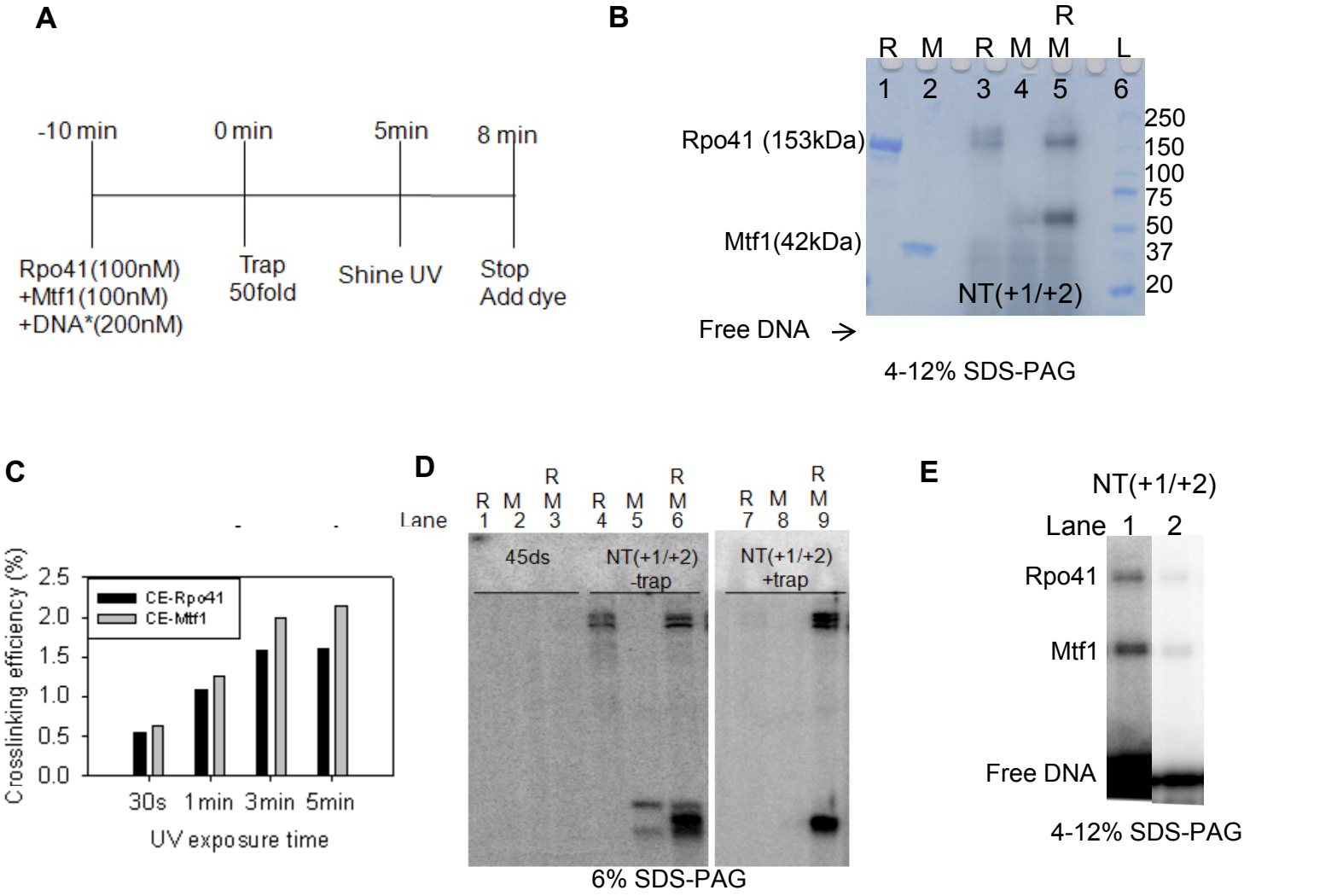

A
-10 min
0 min
5min
8 min
Rpo41(100nM)
+Mtf1(100nM)
+DNA*(200nM)
Trap
50fold
Shine UV
Stop
Add dye
B
R
R M R M M L
1 2 3 4 5 6
Rpo41 (153kDa)
Mtf1(42kDa)
NT(+1/+2)
Free DNA
250
150
100
75
50
37
20
4-12% SDS-PAG
C
Crosslinking efficiency (%)
CE-Rpo41
CE-Mtf1
30s 1min 3min 5min
UV exposure time
D
Lane
R M R M R M
1 2 3 4 5 6 7 8 9
45ds
NT(+1/+2)
-trap
NT(+1/+2)
+trap
6% SDS-PAG
E
NT(+1/+2)
Lane 1 2
Rpo41
Mtf1
Free DNA
4-12% SDS-PAG
F
%intensity of Rpo41
(Intensity of R in R only reaction/Intensity of R in RM reaction)*100
NT(-9/-10)
NT(-5/6)
NT(-4/-5)
NT(-3/-4)
NT(-2/-3)
NT(-1/-2)
NT (+1/+2)
T(-14/-15)
T(-9/-10)
T(-8/-9)
T(-7/-8)
T(-4/-5)
T(-3/-4)
T(-2/-3)
T(-1/-2)
T (+1/+2)
T(+4/+5)
%Intensity of Mtf1
(Intensity of M in M only reaction/Intensity of M in RM reaction)*100
G
Comparison of Rpo41 alone and Rpo41 in complex with Mtf1 on bubble substrates
CE-Rpo41 (%)
CE-Rpo41 in complex
CE-Rpo41 alone
T(-4/-5)B
T(-3/-4)B
T(+1/+2)B
T(+3/+4)B
T(+4/+5)B
NT(-2/-3)JB

Figure S4: Establishing the protein-DNA photo-crosslinking conditions and controls

A) Crosslinking reactions were carried out using 100 nM protein and 300 nM radiolabeled DNA (derivatized at a single defined site). Crosslinked protein-DNA complexes were resolved on a 4-15% SDS-PAGE gel. B) 4-12% SDS-PAGE gel image shows uncrosslinked Rpo41 (lane 1), uncrosslinked Mtf1 (lane 2), Rpo41-DNA (-trap, lane 3), Mtf1-DNA (-trap, lane 4), and Rpo41-Mtf1-DNA (+trap, lane 5) crosslinked to NT(+1/+2) and protein ladder (lane 6). C) Plot shows crosslinking efficiency (as described in Experimental methods) of Rpo41 (black bars) and Mtf1 (grey bars) as a function of time of UV exposure D) Gel image shows the proteins crosslinked to radiolabeled DNA from reactions of Rpo41 (R), Mtf1 (M), or Rpo41-Mtf1 (RM) with underivatized DNA (lanes 1-3), NT(+1/+2) (-trap, lanes 4-6), or NT(+1/+2) in the presence of 50 fold DNA trap (+trap, lanes 7-9). No crosslinked products were observed with underivatized radiolabeled promoter DNA (lanes 1, 2, 3 respectively**)**. Rpo41 alone or Mtf1 alone crosslinked to derivatized DNA (lanes 4, 5), but the crosslinked complexes were reduced in the presence of trap (lanes 7, 8). Crosslinked products of Rpo4-Mtf1 reaction with derivatized DNA prevailed in the presence of the trap. E) Lane 11 from Figure 2b (-ATP panel) is shown (high sensitivity, lane 1 and low sensitivity, lane 2) along with free DNA to show resolution of crosslinked complexes from the free DNA. F) Control crosslinking reactions with Rpo41 alone or Mtf1 alone were analyzed. The plot shows % crosslinking of Rpo41 in an Rpo41-only reaction as compared to the Rpo41-Mtf1 reaction on the various photoactivable DNAs, calculated using the formula: Intensity R in R-only reaction * 100 / Intensity of R in RM reaction. Below is a similar plot for Mtf1 calculated using the formula: Intensity of M in M only reaction*100 / Intensity of M in RM reaction. G) Crosslinking reactions were carried out with Rpo41 (100 nM) and 300 nM radiolabeled premelted promoter DNA (derivatized at a single defined site) in the absence of ATP. The % crosslinking efficiency (CE) for Rpo41 at different positions was calculated using Equation 1 (main text Experimental Procedures). Plot shows crosslinking efficiency of Rpo41 when present alone (grey bars) and when present in complex with Mtf1 (black bars). The black bars (CE-Rpo41) obtained from experiment shown in Figure 3c.

Figure S5

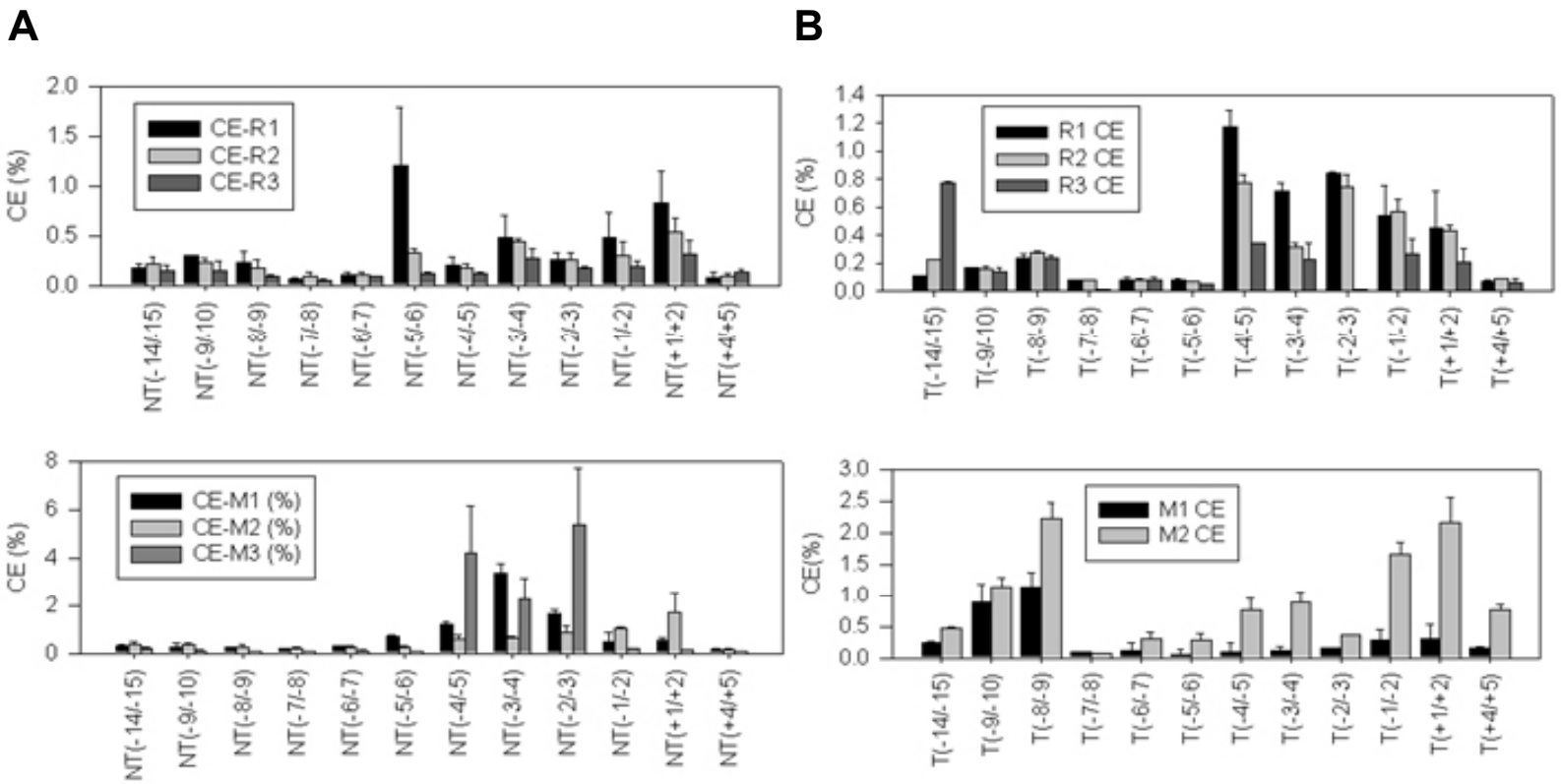


Figure S5: Analysis of the individual bands of Rpo41 and Mtf1 crosslinked to the NT and T strands

A) Crosslinking efficiency (CE) calculations for individual Rpo41 (R1, R2 and R3) and Mtf1 bands (M1, M2 and M3) on Non-template strand.

B) Crosslinking efficiency (CE) calculations for individual Rpo41 (R1, R2 and R3) and Mtf1 bands (M1 and M2) on the Template strand.

Figure S6

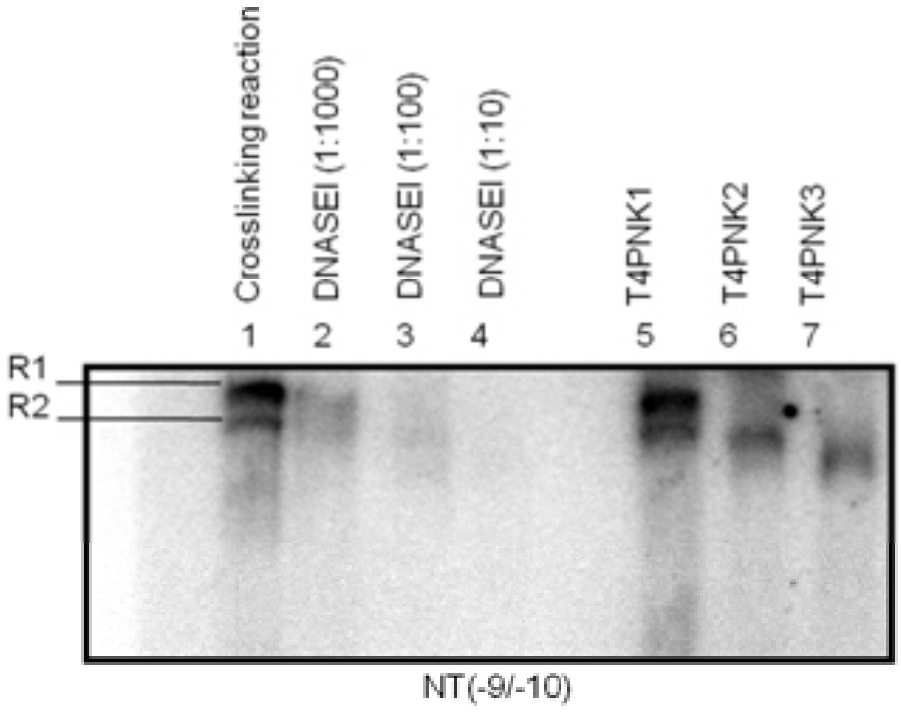


Figure S6: Effect of Dnase I treatment on the mobility of Rpo41 crosslinked complex on SDS-PAGE gel

The crosslinking reaction products of Rpo41-Mtf1-[NT(-9/-10) (as described) were mixed with Dnase I (different dilutions for 30 min. The 5'end of the digested DNA was then relabeled using γ-^{32}P-ATP and T4-PNKase, and analyzed by SDS-PAGE. The 6.5% gel shows the crosslinked Rpo41 bands, R1 and R2 (Lane labeled crosslinking reaction). Lanes 2, 3, 4 show the disappearance of the crosslinked complex in the presence of 1000 fold, 100fold, and 10 fold dilutions of DNASEI respectively. Lanes 5, 6, and 7 show the Rpo41 crosslinked DNA after the relabeling reaction of the DNASEI reactions corresponding to lanes 2, 3 and 4 respectively.

Figure S7

Position	Size of the Rpo41 sphere *	Size of the Mtf1 sphere *
NT(-14/-15)	0.0220	0.0330
NT(-9/-10)	0.0271	0.0269
NT(-8/-9)	0.0198	0.0223
NT(-7/-8)	8.5698e-3	0.0167
NT(-6/-7)	0.0118	0.0240
NT(-5/-6)	0.0659	0.0372
NT(-4/-5)	0.0199	0.2311
NT(-3/-4)	0.0477	0.2165
NT(-2/-3)	0.0273	0.3000
NT(-1/-2)	0.0388	0.0616
NT(+1/+2)	0.0681	0.0902
NT(+4/+5)	0.0125	0.0153
T(-14/-15)	0.0447	0.0369
T(-9/-10)	0.0191	0.1009
T(-8/-9)	0.0303	0.1728
T(-7/-8)	0.0163	0.0467
T(-6/-7)	9.7009e-3	0.0228
T(-5/-6)	8.0891e-3	0.0182
T(-4/-5)	0.0916	0.0440
T(-3/-4)	0.0505	0.0489
T(-2/-3)	0.0269	0.0941
T(-1/-2)	0.0554	0.0936
T(+1/+2)	0.0435	0.1136
T(+4/+5)	8.9631e-3	0.0424
T(+3/+4)B	0.0220	0.0330
T(+4/+5)B	0.0271	0.0269

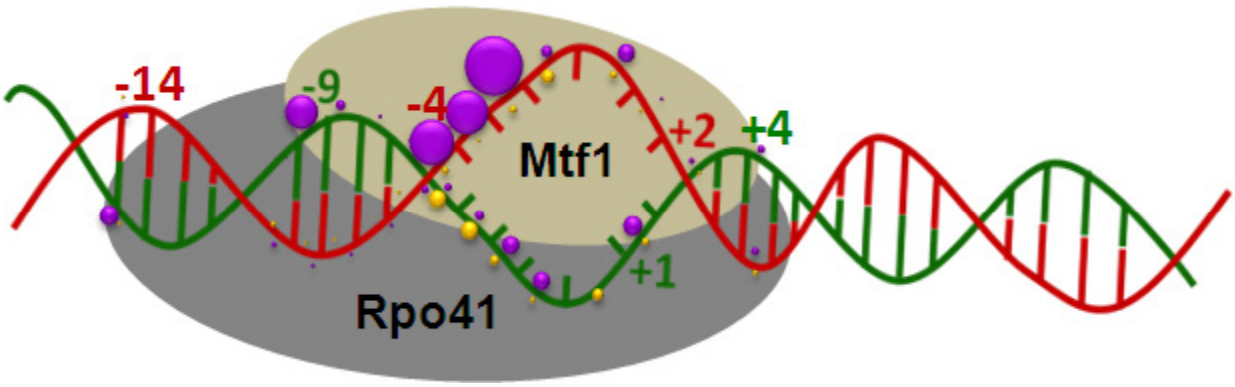


Figure S7: Sphere size calculation based on crosslinking efficiency of Rpo41 and Mtf1 on the NT and T strand

The crosslinking efficiency at each position (from Fig 2) was normalized to the highest crosslinking efficiency seen at NT(-2/-3) for Mtf1 and sphere diameters were calculated according to this normalization. The table shows the sphere diameters for Rpo41 and Mtf1 as normalized to the highest diameter (0.3cm). These spheres were then drawn on the model shown in Fig 4a. The model (also in Fig 4a) is shown below the table.

Figure S8

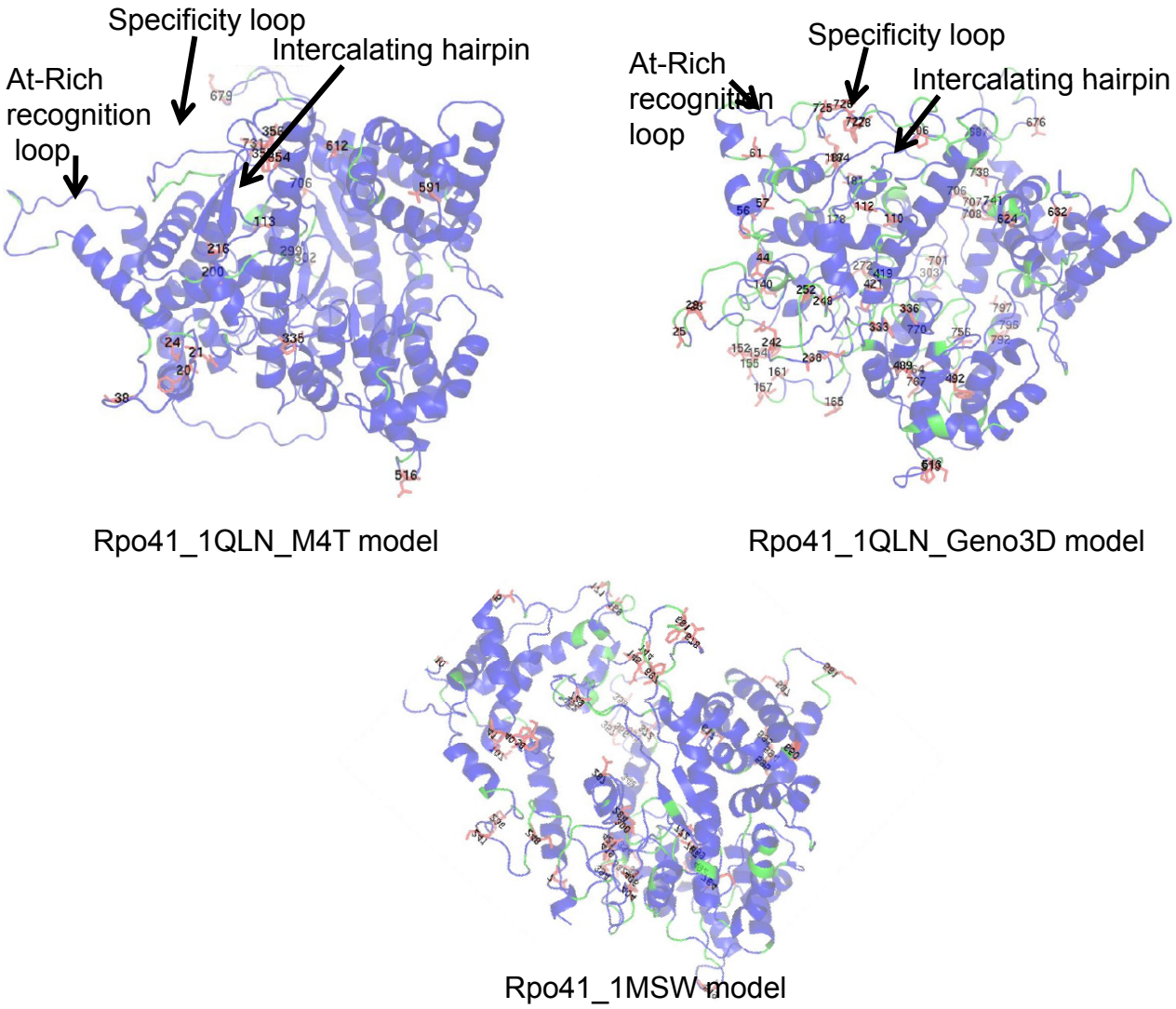


Figure S8: Residue plot of Rpo41_1QLN_M4T model and Rpo41_1QLN_Geno3D and Rpo41_1MSW_Geno3D model model based on Ramachandran plot analysis (Richardson Lab's Molprobity, method B). The Rpo41 model is shown in blue. Residues that are numbered are present in the disallowed regions of the Ramachandran plot.

Figure S9

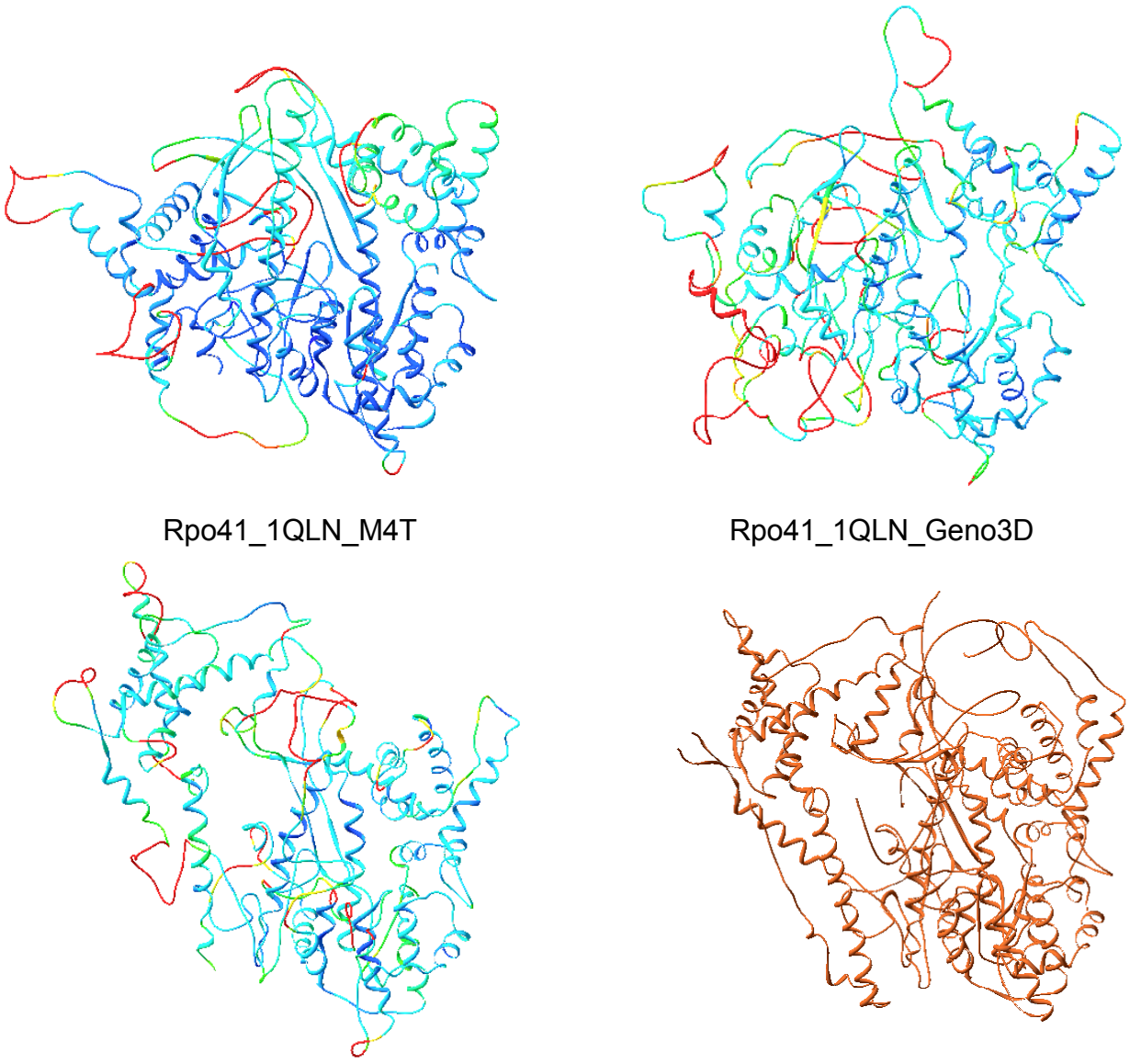
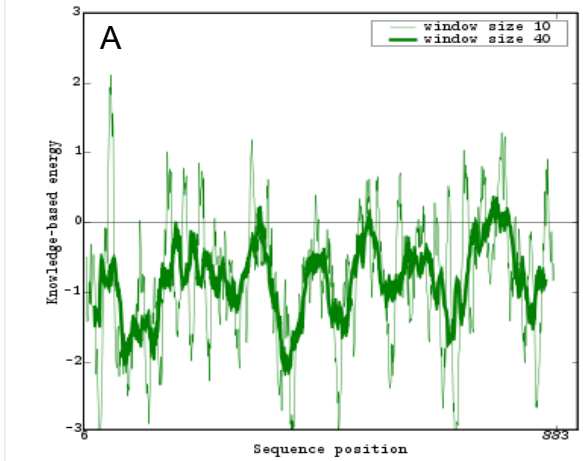
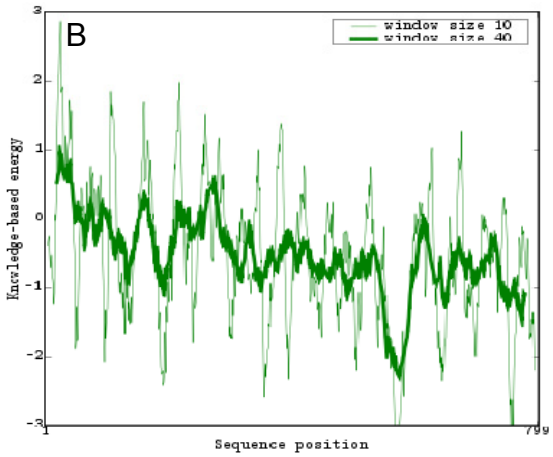


Figure S9: Deviations of the structural models of Rpo41 from template structures
Structural models of Rpo41 in the initiation complex (Rpo41_1QLN_M4T and Rpo41_1QLN_Geno3D) and elongation complex (Rpo41_1MSW_Geno3D) visualized in Swiss-PDB viewer and colored according to deviation in the template structure (1QLN). Blue to red transition represents least to most deviation. Maximum deviation observed was 0.96A in both models. The template structure of T7RNAP in elongation conformation (PDB ID: 1MSW) is provided for reference.

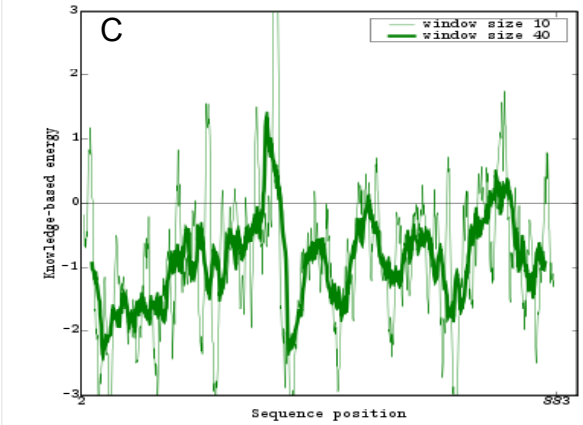
Figure S10



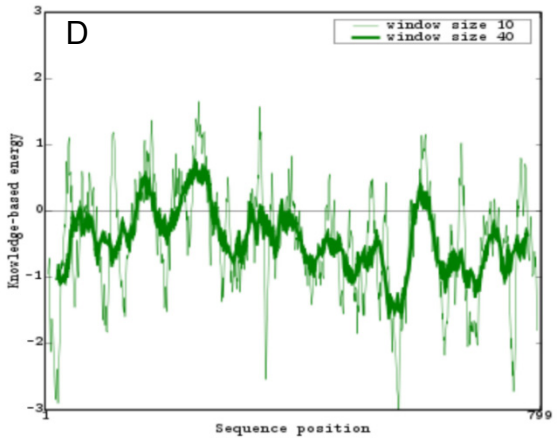
1QLN (*Z*-Score: -11.71)



Rpo41_1QLN model (M4T) *Z*-Score: -9.18



T7RNAP, PDB ID:1MSW (*Z*-Score: -12)



Rpo41_1MSW model (Geno3D)*Z*-Score: -8.64

Figure S10: PROSA analysis for Rpo41_1QLN_M4T model and Rpo41_1MSW_Geno3D model:
Averaged PROSA generated energy scores for (A) the initiation template T7RNAP (1QLN), and (B) one of the Rpo41 models (Rpo41_1QLN_M4T) and also for (C)the elongation template T7RNAP (1MSW), and (D) Rpo41 models(Rpo41_1MSW). plotted against sequence position. Since a plot of single residue energies results in a lot of fluctuations, the data is smoothed by calculating the average energy over a 40-residue interval (thick green line). A second line with a smaller window size of 10 residues is shown in the background of the plot (thin line).

Figure S11

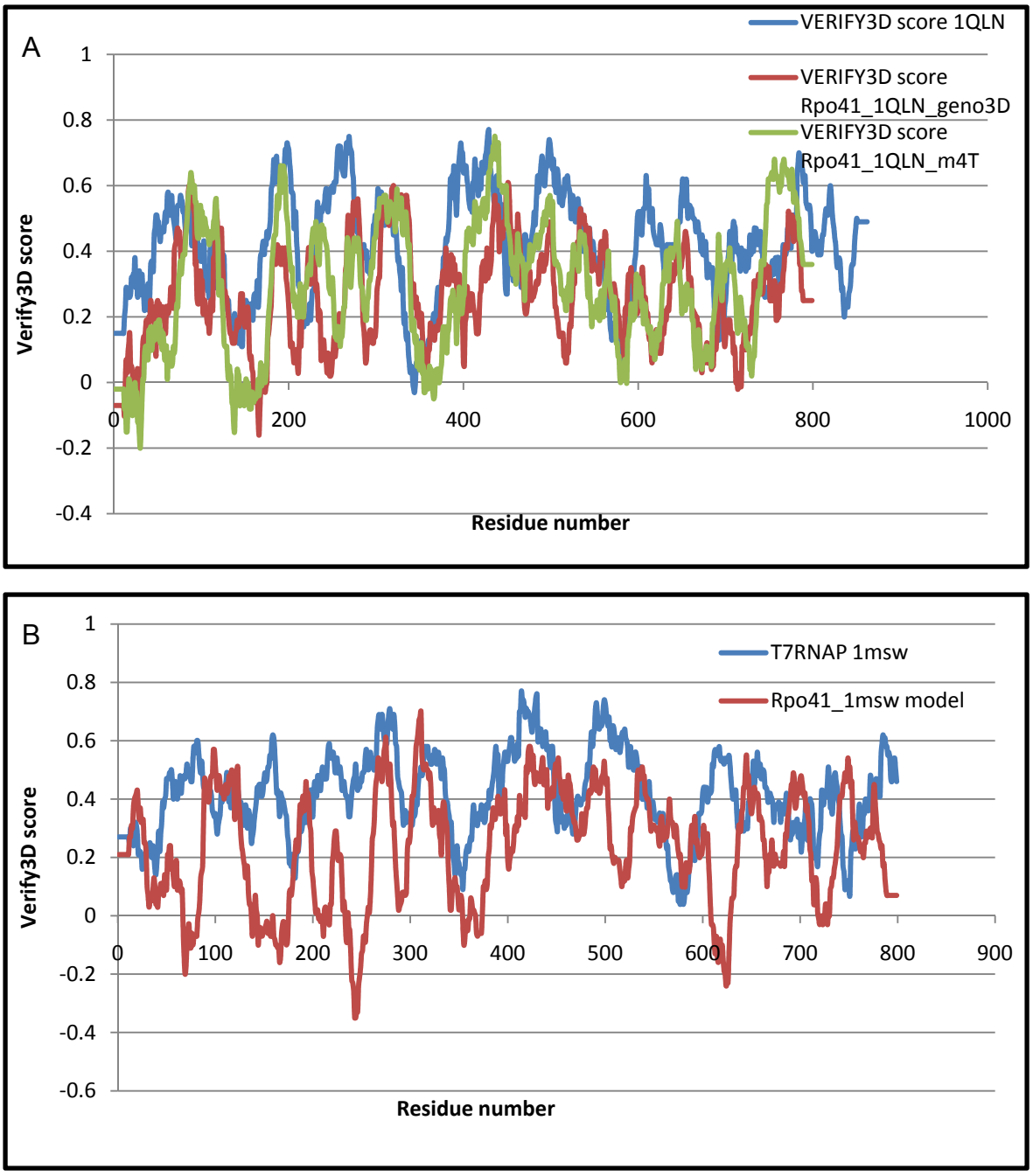


Figure S11: Verify3D scores for Rpo41 models

(A) Verify3d scores for Rpo41 generated models for the initiation conformation. Average 3D-ID scores for each residue in the T7RNAP template, 1QLN (blue), Rpo41_1QLN_Geno3D model (red) and Rpo41_1QLN_M4T model (green) are shown. (B) Verify3d scores for Rpo41 generated models for the elongation conformation. Average 3D-ID scores for each residue in the T7RNAP template, 1MSW (blue), Rpo41_1MSW_Geno3D model (red).

Scores range from -1 (bad) to +1 (good). The starting residue (residue notation 1) in T7RNAP refers to residue number 37 and the starting residue (residue notation 1) in Rpo41 refers to residue number 416. Both models show similar distribution of reasonable 3D-1D scores validating the reliability of the model.

Structural model of Rpo41: Residues 416-1214 of Rpo41 were used as the input sequence for 2 different modeling servers Geno3D (1) and M4T (2,3). Geno3D is an automated program which allows for template selection based on the sequence alignment using ClustalX alignment, and model building using Modeller. For the query sequence (Rpo41 residues 416-1214) Geno3D generated a list of homologous proteins with known 3D structures. Template hits included T7RNAP structures in its initiation (1QLN) and elongation conformations (1MSW). Among the different templates generated, 1QLN was selected as the target structure modeling Rpo41 in the initiation conformation (IC) and 1MSW was selected to model Rpo41 in the elongation conformation (EC). Based on these templates, a 3D model of Rpo41 (for IC and EC) was constructed by Geno3D. M4T is an automated protocol specially designed to build 3D models of proteins which have a low sequence similarity to the template proteins. M4T server performs automated template search and selection, target sequence to template structure alignment using specialized protocols and model building using Modeller. For the query sequence (Rpo41 residues 416-1214), M4T automatically picked 1QLN as the template and built a 3D model based on this selection. Since M4T does not allow user-picked template selection, only Rpo41 model in IC could be achieved using this server.

Residues 416-1214 of Rpo41 show 26% sequence identity to T7RNAP residues (31-816). By both servers, 467 out of 744 residues of Rpo41 show at least semi-conservative alignment to T7RNAP, which accounts for 63% of Rpo41 residues in the selected region for modeling. The alignments generated by M4T and Geno3D are shown below.

M4T generated sequence alignment for Rpo41_1QLN

```
1QLNA                QLALEHESYEMGEARFRKMFPLITT--------------------LLPKMIARINDWFEE   91
Rpo41                QKVLENRATEAARERWKHDFEEAKARGDISIEKNLNVKLWKWYNEMLPLVKEEINHCRSL  474
                     * .**:.: * .. *::: *   .:                    :** :  .**.  .

1QLNA                VKAK----------RGKRPTAFQFLQEIKPEAVAYITIKTTLACLTSA---DNTTVQAVA   138
Rpo41                LSEKLSDKKGLNKVDTNRLGYGPYLTLIDPGKMCVITILELLKLNSTGGVIEGMRTARAV   535
                     :. *           :*     :*  *.*  :. ***   *   ::.   :.  .  ..

1QLNA                SAIGRAIEDEARFGRIRDLEAKHFKKNVEEQLNKRVGHVYKKAFMQVVEADMLSKGLLGG   198
Rpo41                ISVGKAIEMEFRSEQVLKSESQAFRDVNKKS------PEFKKLVQN---AKSVFRSSQIE   576
                      ::*:*** * *  :: . *:: *:.  ::.        :** . :   *. : :.

1QLNA                EAWSSWHKEDSIHVGVRCIEMLIESTGM-----------------VSLHRQNAGVVGQDS   241
Rpo41                QSKILWPQSIRARIGSVLISMLIQVAKVSVQGVDPVTKAKVHGEAPAFAHGYQYHNGSKL   646
                     ::   * :.   ::*   *.***: : :                   :: :      *..

1QLNA                ETIELAPEYAEAIATRAGALAGISPMFQPCVVPPKPWTGITGGGYWANGRRPLALVRTHS   301
Rpo41                GVLKIHKTLIRQLNGE-RLIASVQPQLLPMLVEPKPWVNWRSGGYHYTQSTLLRTKDSPE   705
                      .:::     . :  .   :*.:.* : * :* ****..  .***  .    *    : .

1QLNA                KK-ALMR-YEDVYMPEVYKAINIAQNTAWKINKKVLAVANVITKWKHCPVEDIPAIEREE   359
Rpo41                QVAYLKAASDNGDIDRVYDGLNVLGRTPWTVNRKVFDVVSQVWN-KGEGFLDIPGAQDEM   764
                     :   *    ::  : .**..:*:  .*.*.:*:**: *.. : : *   . ***. : *

1QLNA                LPMKPEDIDMNPEALTAWKRAAAAVYRKDKARKSRRISLEFMLEQANKFANHKAIWFPYN   419
Rpo41                VLPPAPPKNSDPSILRAWKLQVKTIANKFSSDRSNRCDTNYKLEIARAFLGE-KLYFPHN   823
                     :   .   : :*. * ***  . :: .* .: :*.* . :: ** *. * ..  ::**:*

1QLNA                MDWRGRVYAV-SMFNPQGNDMTKGLLTLAKGKPIGKEGYYWLKIHGANCAGVDKVPFPER
Rpo41                LDFRGRAYPLSPHFNHLGNDMSRGLLIFWHGKKLGPSGLKWLKIHLSNLFGFDKLPLKDR   883
                     :*:***.*.: . **  ****::*** : :** :* .*  ***** :*  *.**:*: :*

1QLNA                IKFIEENHENIMACAKSPLEN-TWWAEQDSPFCFLAFCFEYAGVQHH--GLSYNCSLPLA   478
Rpo41                VAFTESHLQDIKDSAENPLTGDRWWTTADKPWQALATCFELNEVMKMDNPEEFISHQPVH   943
                     : * *.: ::*  .*:.** .  **:  *.*:  ** ***   * :     .: .  *:

1QLNA                FDGSCSGIQHFSAMLRDEVGGRAVNLLPSETVQDIYGIVAKKVNEILQADAINGTDNEVV   595
Rpo41                QDGTCNGLQHYAALGGDVEGATQVNLVPSDKPQDVYAHVARLVQKRLEIAAEKGDE----   999
                      **:*.*:**::*:  *  *.  ***:**:. **:*. **: *:: *:  * :* :

1QLNA                TVTDENTGEISEKVKLGTKALAGQWLAYGVTRSVTKRSVMTLAYGSKEFGFRQQVLEDTI   655
Rpo41                ----------------------NAKILKDKITRKVVKQTVMTNVYGVTYVGATFQIAKQLS  1038
                                        .: *   :**.*.*::*** .** . .*   *: ::

1QLNA                QPAIDSGKGLMFTQPNQAAGYMAKLIWESVSVTVVAAVEAMNWLKSAAKLLAAEVKDKK-   714
Rpo41                PIFDDR------KESLDFSKYLTKHVFSAIRELFHSAHLIQDWLGESAKRISKSIRLDVD  1092
                         *       .:. : : *::* ::.::   . :*    :** .:** :: .:: .

1QLNA                -----TGEILRKRCAVHWVTPDGFPVWQEYKKPIQTRLNLMFLGQFRLQPTINTNKDSEI   769
Rpo41                EKSFKNGNKPDFMSSVIWTTPLGLPIVQPYREESKKQVETNLQTVFI----SDPFAVNPV  1148
                          .*:     .:* *.** *:*: * *::  :.:::  :   *      :.   . :


1QLNA                DAHKQESGIAPNFVHSQDGSHLRKTVVWAHEKYGIESFALIHDSFGTIPADAANLFKAVR   829
Rpo41                NARRQKAGLPPNFIHSLDASHMLLSAAECGK-QGL-DFASVHDSYWTHASDIDTMNVVLR  1206
                     :*::*::*:.***:** *.**:  :.. . :  *: .** :***: * .:*  .:  .:*

1QLNA                ETMVDTYESCDVLADFYDQFADQLHESQLDKMPALPAKGNLNLRDILESDFAFA         883
Rpo41                EQFIKLHE-----------------------------------------------       1214
                     * ::. :*
```

Geno3D generated sequence alignment for Rpo41_1QLN

```
CLUSTAL W(1.81) multiple sequence alignment


pdb1qlnA_0            QLALEHESYEMGEARFRKMFXXXXXXXXXXXXXXXX----PLITTLLPKMIARIND----    87
Rpo41x0_0             QKVLENRATEAARERWKHDFEEAKARGDISIEKNLNVKLWKWYNEMLPLVKEEINHCRSL   475
                      * .**:.: * .. *::: *                       . :** :  .**.

pdb1qlnA_0            WFEEVKAKRG------KRPTAFQFLQEIKPEAVAYITIKTTLACLTSAD----NTTVQAV   147
Rpo41x0_0             LSEKLSDKKGLNKVDTNRLGYGPYLTLIDPGKMCVITILELLKLNSTGGVIEGMRTARAV   535
                        *::. *:*      :*     :*  *.*  :. ***   *   ::..      *.:**

pdb1qlnA_0            ASAIGRAIEDEARFGRIRDLEAKHF----KKNVE-EQLNKRVGHVYKKAFMQVVEADMLS   192
Rpo41x0_0             IS-VGKAIEMEFRSEQVLKSESQAFRDVNKKSPEFKKLVQNAKSVFRSS--QIEQSKIL-   591
                       * :*:*** * *  :: . *:: *    **. * ::* :..  *::.:  *: ::.:*

pdb1qlnA_0            KGLLGGEAWSSWHKEDSIHVGVRCIEMLIESTGMVSLHRQNAGVVGQDSETIELAPEYAE   252
Rpo41x0_0             -----------WPQSIRARIGSVLISMLIQVAKVSVQGVDPVTKAKVHGEAPAFAHGYQY   640
                                * :.   ::*   *.***: : :     : .  .  ..*:  :*  *

pdb1qlnA_0            AIATRAGAL-----------------AGISPMFQPCVVPPKPWTGITGGGYWANGRRPLAL   296
Rpo41x0_0             HNGSKLGVLKIHKTLIRQLNGERLIASVQPQLLPMLVEPKPWVNWRSGGYHYTQS---TL   697
                        .:: *.*               *.:.* : * :* ****..  .***  .     :*

pdb1qlnA_0            VRTHSKKALMRY-----EDVYMPEVYKAINIAQNTAWKINKKVLAVANVITKW-KHCPVE   350
Rpo41x0_0             LRTKDSPEQVAYLKAASDNGDIDRVYDGLNVLGRTPWTVNRKVFDVVSQV--WNKGEGFL   755
                      :**:..   : *     ::  : .**..:*:  .*.*.:*:**: *.. :  * *   .

pdb1qlnA_0            DIPAIEREEL--PMKPEDIDMNPEALTAWKRAAAAVYRKDKARKSRRISLEFMLEQANKF   408
Rpo41x0_0             DIPGAQDEMVLPPAPPKNSD--PSILRAWKLQVKTIANKFSSDRSNRCDTNYKLEIARAF   815
                      ***. : * :  *  *:: *  *. * ***  . :: .* .: :*.* . :: ** *. *

pdb1qlnA_0            ANHKAIWFPYNMDWRGRVYAVS-MFNPQGNDMTKGLLTLAKGKPIGKEGYYWLKIHGANC   467
Rpo41x0_0             LGEK-LYFPHNLDFRGRAYPLSPHFNHLGNDMSRGLLIFWHGKKLGPSGLKWLKIHLSNL   874
                       ..* ::**:*:*:***.*.:*  **  ****::*** : :** :* .*  ***** :*

pdb1qlnA_0            AGVDKVPFPERIKFIEENHENIMACAKSPLE-NTWWAEQDSPFCFLAFCFEYAGVQH--H   524
Rpo41x0_0             FGFDKLPLKDRVAFTESHLQDIKDSAENPLTGDRWWTTADKPWQALATCFELNEVMKMDN   936
                       *.**:*: :*: * *.: ::*  .*:.**  : **:  *.*:  ** ***   * :  :

pdb1qlnA_0            GLSYNCSLPLAFDGSCSGIQHFSAMLRDEVGGRAVNLLPSETVQDIYGIVAKKVNEILQA   584
Rpo41x0_0             PEEFISHQPVHQDGTCNGLQHYAALGGDVEGATQVNLVPSDKPQDVYAHVARLVQKRLEI   934
                        .: .  *:  **:*.*:**::*:  *  *.  ***:**:. **:*. **: *:: *:

pdb1qlnA_0            DAINGTDNEVVTVTDENTGEISEKVKLGTKALAGQWLAYGVTRSVTKRSVMTLAYGSKEF   644
Rpo41x0_0             AAEKGDEN-------------------------AKILKDKITRKVVKQTVMTNVYGVTYV   994
                       * :* :*                         .: *   :**.*.*::*** .** . .

pdb1qlnA_0            GFRQQVLEDTIQPAIDSGK-GLMFTQPNQAAGYMAKLIWESVSVTVVAAVEAMNWLKSAA   703
Rpo41x0_0             GATFQIAKQ-LSPIFDDRKESLDFSK------YLTKHVFSAIRELFHSAHLIQDWLGESA  1028
                      *   *: :: :.* :*. * .* *::      *::* ::.::   . :*    :** .:*

pdb1qlnA_0            KLLAAEV------KDKKTGEILRKRCAVHWVTPDGFPVWQEY----KKPIQTRLNLMFLG   753
Rpo41x0_0             KRISKSIRLDVDEKSFKNGNKPDFMSSVIWTTPLGLPIVQPYREESKKQVETNLQTVFIS  1092
                      * :: .:      *. *.*:     .:* *.** *:*: * *    ** ::*.*: :*:.

pdb1qlnA_0            Q-FRLQPTINTNKDSEIDAHKQESGIAPNFVHSQDGSHLRKTVVWAHE--KYGIESFALI   810
Rpo41x0_0             DPFAVNP----------VNARRQKAGLPPNFIHSLDASHM---LLSAAECGKQGLD-FASV  1178
                      : * ::*         ::*::*::*:.***:** *.**:   :: * *  * *:: ** :

pdb1qlnA_0            HDSFGTIPADAANLFKAVRETMVDTYE                                    883
Rpo41x0_0             HDSYWTHASDIDTMNVVLREQFIKLHE                                   1214
                      ***: * .:*  .:  .:** ::. :*
```

Geno3D generated sequence alignment for Rpo41_1MSW

```
CLUSTAL W(1.81) multiple sequence alignment


pdb1mswD_0            QLALEHESYEMGEARFRKMFERQLKAGEVAD----NAAAKPLITTLLPKMIARIND----    97
Rpo41x0_0             QKVLENRATEAARERWKHDFEEAKARGDISIEKNLNVKLWKWYNEMLPLVKEEINHCRSL   475
                      * .**:.: * .. *::: **.    *:::     *.      . :** :  .**.

pdb1mswD_0            WFEEVKAKRG------KRPTAFQFLQEIKPEAVAYITIKTTLACLTSAD----NTTVQAV   148
Rpo41x0_0             LSEKLSDKKGLNKVDTNRLGYGPYLTLIDPGKMCVITILELLKLNSTGGVIEGMRTARAV   535
                        *::. *:*      :*     :*  *.*  :. ***   *   ::..      *.:**

pdb1mswD_0            ASAIGRAIEDEARFGRIRDLEAKHF----KKNVE-EQLNKRVGHVYKKAFMQVVEADMLS   192
Rpo41x0_0             IS-VGKAIEMEFRSEQVLKSESQAFRDVNKKSPEFKKLVQNAKSVFRSS--QIEQSKIL-   591
                       * :*:*** * *  :: . *:: *    **. * ::* :..  *::.:  *: ::.:*

pdb1mswD_0            KGLLGGEAWSSWHKEDSIHVGVRCIEMLIESTGMVSLHRQXXXXXXXXSETIELAPEYAE   252
Rpo41x0_0             -----------WPQSIRARIGSVLISMLIQVAKVSVQGVDPVTKAKVHGEAPAFAHGYQY   640
                                * :.   ::*   *.***: : :     :        .*:  :*  *

pdb1mswD_0            AIATRAGAL----------------AGISPMFQPCVVPPKPWTGITGGGYWANGRRPLAL   286
Rpo41x0_0             HNGSKLGVLKIHKTLIRQLNGERLIASVQPQLLPMLVEPKPWVNWRSGGYHYTQS---TL   697
                        .:: *.*               *.:.* : * :* ****..  .***  .     :*

pdb1mswD_0            VRTHSKKALMRY-----EDVYMPEVYKAINIAQNTAWKINKKVLAVANVITKW-KHCPVE   350
Rpo41x0_0             LRTKDSPEQVAYLKAASDNGDIDRVYDGLNVLGRTPWTVNRKVFDVVSQV--WNKGEGFL   755
                      :**:..   : *     ::  : .**..:*:  .*.*.:*:**: *.. :  * *   .

pdb1mswD_0            DIPAIEREELPMKXXXXXXXXXXXTAWKRAAAAVYRKDKARKSRRISLEFMLEQANKFAN   408
Rpo41x0_0             DIPGAQDEMVLPPAPPKNSDPSILRAWKLQVKTIANKFSSDRSNRCDTNYKLEIARAFLG   815
                      ***. : * :               ***  . :: .* .: :*.* . :: ** *. * .

pdb1mswD_0            HKAIWFPYNMDWRGRVYAVS-MFNPQGNDMTKGLLTLAKGKPIGKEGYYWLKIHGANCAG   467
Rpo41x0_0             EK-LYFPHNLDFRGRAYPLSPHFNHLGNDMSRGLLIFWHGKKLGPSGLKWLKIHLSNLFG   874
                      .* ::**:*:*:***.*.:*  **  ****::*** : :** :* .*  ***** :*  *

pdb1mswD_0            VDKVPFPERIKFIEENHENIMACAKSPLE-NTWWAEQDSPFCFLAFCFEYAGVQH--HGL   524
Rpo41x0_0             FDKLPLKDRVAFTESHLQDIKDSAENPLTGDRWWTTADKPWQALATCFELNEVMKMDNPE   934
                      .**:*: :*: * *.: ::*  .*:.**  : **:  *.*:  ** ***   * :  :

pdb1mswD_0            SYNCSLPLAFDGSCSGIQHFSAMLRDEVGGRAVNLLPSETVQDIYGIVAKKVNEILQADA   594
Rpo41x0_0             EFISHQPVHQDGTCNGLQHYAALGGDVEGATQVNLVPSDKPQDVYAHVARLVQKRLEIAA   994
                      .: .  *:  **:*.*:**::*:  *  *.  ***:**:. **:*. **: *:: *:  *

pdb1mswD_0            INGTDNEVVTVTDENTGEISEKVKLGTKALAGQWLAYGVTRSVTKRSVMTLAYGSKEFGF   644
Rpo41x0_0             EKGDEN-------------------------AKILKDKITRKVVKQTVMTNVYGVTYVGA  1029
                       :* :*                          .: *   :**.*.*::*** .** . .*

pdb1mswD_0            RQQVLEDTIQPAIDSGK-GLMFTQPNQAAGYMAKLIWESVSVTVVAAVEAMNWLKSAAKL   703
Rpo41x0_0             TFQIAKQ-LSPIFDDRKESLDFSK------YLTKHVFSAIRELFHSAHLIQDWLGESAKR  1082
                        *: :: :.* :*. * .* *::      *::* ::.::   . :*    :** .:**

pdb1mswD_0            LAAEV------KDKKTGEILRKRCAVHWVTPDGFPVWQEY----KKPIQTRLNLMFLGQ-   753
Rpo41x0_0             ISKSIRLDVDEKSFKNGNKPDFMSSVIWTTPLGLPIVQPYREESKKQVETNLQTVFISDP  1142
                      :: .:      *. *.*:     .:* *.** *:*: * *    ** ::*.*: :*:.:

pdb1mswD_0            FRLQPTINTNKDSEIDAHKQESGIAPNFVHSQDGSHLRKTVVWAHE--KYGIESFALIHD   910
Rpo41x0_0             FAVNP---------VNARRQKAGLPPNFIHSLDASHM---LLSAAECGKQGLD-FASVHD  1179
                      * ::*         ::*::*::*:.***:** *.**:   :: * *  * *:: ** :**

pdb1mswD_0            SFGTIPADAANLFKAVRETMVDTYE                                      883
Rpo41x0_0             SYWTHASDIDTMNVVLREQFIKLHE                                     1214
                      *: * .:*  .:  .:** ::. :*
```

Assessment of the quality of the model: The quality of the obtained models was evaluated using PROCHECK(4), MOLPROBITY (5), WHATIF(6), PROSA(7), VERIFY3D(8) and ERRAT(9). The secondary structure predictions were performed using PSIPred(10), PORTER (11)

Ramachandran plot: The geometry of the model was evaluated using the Ramachandran plot calculations in PROCHECK. This method checks for the stereochemical correctness of the modeled amino acids in the chain. Stereochemical evaluations of the backbone Psi and Phi dihedral angles suggest that in each model more than 95% of the residues fall in the most favored and the additional allowed regions of the Ramachandran plots. Of interest to this study, none of the residues in the disallowed regions are present within or in the vicinity of the predicted AT-Rich loop, specificity loop, or the intercalating hairpin regions of Rpo41 (refer manuscript). The Table below shows that the model generated by M4T server is of a higher quality than the Geno3D generated model. However, the model quality of the predicted DNA binding elements is similar in both M4T and Geno3D generated models (see below).

Ramachandran plot regions	Rpo41_1QLN_M4T		Rpo41_1QLN_Geno3D		Rpo41_1MSW_Geno3D	
Calculation method	A	B*	A	B*	A	B*
Most favored regions	87.6%	91.3%	60.8%	69.4%	67.2%	76.8%
Allowed regions	11.8%	6.3%	34%	23.6%	29.9%	16.8%
Disallowed regions	0.6%	2.4%	5.2%	7%	2.8%	6.4%

*Residue plots generated from Method B indicate which residues lie in the disallowed regions of the Ramachandran plot. As can be seen in all three models (Figure S8), with the exception of residues numbers 725, 726, 728 (which correspond to actual residue numbers 1140, 1141 and 1143 in Rpo41) none of the residues lie in the predicted DNA-binding elements of Rpo41 (Figure S8).

Method A: Laskowski R A et al "PROCHECK: a program to check the stereochemical quality of protein structures" J Appl Cryst, 26 (1993): 283-291. Method B: Simon C. Lovell, Ian W. Davis, W. Bryan Arendall III, Paul I. W. de Bakker, J. Michael Word, Michael G. Prisant, Jane S. Richardson, David C. Richardson (2003) Structure validation by C-alpha geometry: phi, psi, and C-beta deviation. Proteins: Structure, Function, and Genetics. 50: 437-450.

RMS deviation: Deviations in the Rpo41 models from the template structures were calculated and accordingly colored in Swiss-PDB viewer(12). Blue to red transition represents least to most deviations from the template structure. Maximum deviation observed was 0.96A in both models, and the areas of maximum deviations were either completely away or within loops of the abovementioned predicted DNA binding elements (Figure S9).

WHATIF: The model was also run through WHATIF program (6) to determine if there were any severe warnings with respect to bond angles and bond lengths, or if some residues deviated severely from planarity. Some errors and warnings were observed, but these residues were away from all the region of the proposed elements of DNA binding in the protein.

PROSA: PROSA is a statistical method of checking 3D models based on knowledge-based energy potentials(7). PROSA measures energies of a generated model and measures its deviation from those compiled from the database of experimentally known structures and determines a statistical average. Model quality can be assessed by plotting these energies as a function of amino acid position. Positive values correspond to problematic and erroneous parts of the model. Negative values ensure correctness of the model. From the plots (Figure S10), most of the protein in both models shows negative values, and hence confirms the reliability of the models.

Verify3D: Verify3D (8)is a statistical method of model assessment. It analyzes the compatibility of an atomic model (3D) with its own amino acid sequence (1D). Verify3D assigns an environmental class to each residue of the protein based on the secondary structure, area buried, polar contacts etc. A total of 18 environmental classes are considered. The probability with which each amino acid type is present in each environment is calculated. The sum of these probabilities is scored within a 21 residue window. If the probability is low, then the model is incorrect. The Verify3D score ranges from -1 (bad) to +1 (good). Average 3D-ID scores for each residue in the T7RNAP template, 1QLN (blue), Rpo41_1QLN_M4T model (green) and Rpo41_1QLN_Geno3D model (red) show similar distribution of reasonable 3D-1D scores (Figure S11) validating the reliability of both models.

ERRAT: ERRAT (9) is a method which identifies whether regions in a protein model have been correctly or incorrectly determined. This method primarily analyzes the pairwise non-covalently bonded interaction statistics between the atoms C, N and O (leading to 6 different types of such interactions; CC, CN, CO, NN, NO, OO). It calculates the fraction of all interactions of a particular type for each residue, and produces a quadratic error function. A typical output comprises of a plot of this error function vs position of a 9-residue sliding window. Statistical confidence limits are determined using error values from 96 highly refined structures. A good model should have not more than 5% protein above the 95% confidence limit. Both Rpo41_1QLN models as well as the Rpo41_1MSW model have 96% of the residues below the 95% confidence limit, suggesting the goodness of the model.

After assessment and structural validation, the models of Rpo41 were superimposed onto the structure of T7RNAP (PDB ID: 1QLN) in the program UCSF Chimera using the command MATCHMAKER. T7RNAP (1QLN) was selected as the reference chain and Rpo41 model was selected as the query chain for structure comparison using Smith-Waterman alignment and other default parameters.

Secondary structure predictions in Rpo41 1QLN models: We compared the sequence-predicted secondary structures to the model generated secondary structures of three elements hypothesized to be present in Rpo41 (refer to manuscript), namely the predicted AT-Rich loop, the specificity loop and the intercalating B-hairpin. The secondary structure predictions of the element as well as the flanking residues in both the sequence as well as the model suggest that these elements may be present in Rpo41.

The numbers above the amino acid sequence (AA) show the confidence (Conf) on the scale of 10 with which the secondary structures have been predicted. Higher the score, better the prediction. Secondary structure predictions from PSIPred server and PORTER server are shown. The secondary structures present in the Rpo41_1QLN_M4T model (M4T) and Rpo41_1QLN_Geno3D model (Geno3D) are shown.

Predicted AT-RICH loop (460-510, underlined) and residues around the predicted element using sequence based secondary structure predictions and model based secondary structure assignments.

```
Conf    :999999999998688878999999741621124504776997099999999999
AA      :EMLPLVKEEINHCRSLLSEKLSDKKGLNKVDTNRLGYGPYLTLIDPGKMCVITIL
PSIPred :HHHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCHHHHHHHHCCCHHHHHHHHHH
Porter  :HHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCHHHHHHCCCCHHHHHHHHH
M4T     :HHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCHHHHHCCCCCHHHHHHHH
Geno3D  :CHHHHHHHHHHCCCCCCHHHHHCCCCCCCCCCCCCCCCHHHHHCCCCCCCCCCHHH
```

Predicted SPECIFICITY LOOP (1127-1149, underlined) and residues around the predicted element using sequence based secondary structure predictions and model based secondary structure assignments. (Rpo41 sequence shown from 1105-1173)

```
Conf   :881699889998788877524438997057659997167778787888986058348776889999999
AA     :MSSVIWTTPLGLPIVQPYREESKKQVETNLQTVFISDPFAVNPVNARRQKAGLPPNFIHSLDASHMLLS
PSIPRED:CCCEEEECCCCCEEEEEEEECCEEEEECCCEEEEECCCCCCCCCCHHHHHCCCCHHHHHHHHHHHHH
Porter :CCCEEEECCCCCEEEECCEECCCCCCCCCCECCCCCCCCCCCEECHHHHHHHHHHHHHHHHHHHHHH
M4T    :CCCEEEECCCCCCEEECCEECCCCCCCCCCCCCCCCCCCCEEEEHHHHHHHHHHHHHHHHHHHHHH
Geno3D :CCCCCEECCCCCEECCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHCCCC
```

Predicted Intercalating B-hairpin (617-640, underlined) and residues around the predicted element using sequence based secondary structure predictions and model based secondary structure assignments. (Rpo41 sequence shown from 630-660)

```
Conf       : 027899999788899999758999998821530 2
AA         : PAFAHGYQYHNGSKLGVLKIHKTLIRQLNGERLI
PSIPRED    : EEEEEEEEEECCEEEEEEEECHHHHHHHHCCCHH
Porter     : HHHHHHHHHHHHHHCCHEEECHHHHCCCCCCCCC
M4T        : CEEEEECCCCCCCCEEEEEECHHHHHHHHCCCCC
Geno3D     : CEEECHHHHHHHHCCCCCCCCCCCCCCCCCC
```

1. Combet, C., Jambon, M., Deleage, G., and Geourjon, C. (2002) *Bioinformatics* **18**, 213-214
2. Fernandez-Fuentes, N., Madrid-Aliste, C. J., Rai, B. K., Fajardo, J. E., and Fiser, A. (2007) *Nucleic Acids Res* **35**, W363-368
3. Rykunov, D., Steinberger, E., Madrid-Aliste, C. J., and Fiser, A. (2009) *J Struct Funct Genomics* **10**, 95-99
4. R.A. Laskowski, M. W. M., D.S. Moss and J.M. Thornton. (1993) *J. Appl. Crystallogr.*, 283-291
5. Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S., and Richardson, D. C. (2003) *Proteins* **50**, 437-450
6. Hooft, R. W., Vriend, G., Sander, C., and Abola, E. E. (1996) *Nature* **381**, 272
7. Wiederstein, M., and Sippl, M. J. (2007) *Nucleic Acids Res* **35**, W407-410
8. Eisenberg, D., Luthy, R., and Bowie, J. U. (1997) *Methods Enzymol* **277**, 396-404
9. Colovos, C., and Yeates, T. O. (1993) *Protein Sci* **2**, 1511-1519
10. McGuffin, L. J., Bryson, K., and Jones, D. T. (2000) *Bioinformatics* **16**, 404-405
11. Pollastri, G., and McLysaght, A. (2005) *Bioinformatics* **21**, 1719-1720
12. Kaplan, W., and Littlejohn, T. G. (2001) *Brief Bioinform* **2**, 195-197