Structural basis of directionality control in large serine integrases

Heewhan Shin[1], Ying Pigli[1], Tania Peña Reyes[1], James R. Fuller[1], Femi J. Olorunniji[2]*, Phoebe A. Rice[1]*

[1]Department of Biochemistry & Molecular Biology, The University of Chicago; Chicago IL, 60637, USA.

[2]School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University; Liverpool, L3 3AF, UK

*e-mail for co-corresponding authors: Phoebe A. Rice price@uchicago.edu; Femi J. Olorunniji f.j.olorunniji@ljmu.ac.uk

Abstract

Large serine integrases (LSIs) catalyze unidirectional site-specific DNA recombination reactions, yet those reactions are reversed by the presence of a cognate recombination directionality factor (RDF). Mechanistic understanding of directionality control has been hampered by a lack of structural information. Here, we use cryo-electron microscopy (cryo-EM) to determine the structures of six SPbeta integrase—DNA complexes along the integrative (-RDF) and excisive (+RDF) reaction pathways, at 4.16-7.18Å resolution. Our findings reveal how RDF-mediated repositioning of an integrase subdomain (1) dictates which pairs of DNA sites can be assembled into a synaptic complex to initiate recombination and (2) dictates which product complexes will be conformationally locked, preventing the back reaction. These mechanistic insights provide a conceptual framework for engineering efficient and versatile genome editing tools.

Large serine integrases (LSIs) and their cognate recombination directionality factors (RDFs) mediate site-specific DNA recombination reactions that move phage or mobile genetic element DNAs into and out of host bacterial genomes[1]. Several features of these systems make them useful as genetic tools[2,3]. For integrative recombination, only integrase and an ~50bp phage DNA site (*attP*) and an ~40bp bacterial site (*attB*) are required[4–6]. The products of this unidirectional reaction are two hybrid integrase binding sites flanking the integrated phage DNA, *attL and attR* (Figure 1a). Binding of the RDF to the integrase triggers the reverse (excisive) reaction[5–10]. No DNA repair is needed after either reaction because all broken phosphodiester bonds are religated (Figure 1B). The large number of known integrases (and smaller number of known integrase-RDF pairs) provides an array of sequence specificity options. Finally, new tools for inserting *att* sites into heterologous genomes have streamlined the use of LSIs in genome editing[11–18].

LSIs share a conserved catalytic domain and fundamental mechanism with smaller recombinases from the serine family such as resolvases and invertases[1,19–21]. Within a synaptic complex, which contains a recombinase tetramer and two partner DNAs, a key serine residue from each subunit attacks a DNA phosphate, displacing a 3’ OH and generating double-strand breaks with 2-nt 3’ overhangs. In this intermediate state, each 5’ end is covalently linked to a protein subunit (Figure 1B,C, and S1), and the two halves of the synaptic complex can swivel relative to one another about a central flat, hydrophobic interface to reposition the broken ends[22]. Following a ~180° rotation, the chemical reverse of the cleavage reaction can occur, which religates the DNA in the

recombinant configuration[22–24]. Notably, the forward progression of serine recombinase-catalyzed reactions is not driven by changes in chemical bond energy; rather, it is driven by energetic differences between the conformational features of the protein-bound substrates vs. the resulting products[19]. However, understanding the control of reaction directionality for the LSIs has been a long-standing challenge due to the lack of 3D structures for the relevant protein-DNA complexes.

A key control point for serine recombinase regulation is the dimer – to – tetramer transition. Serine recombinases generally bind their cognate DNA sites as catalytically inactive dimers. Tetramerization not only synapses two partner DNAs but also entails a conformational change within the catalytic domains that activates them. However, in the absence of additional tetramer-stabilizing factors, the dimer conformation is strongly favored[25–27]. The additional factors that control this conformational equilibrium vary widely among different subfamilies of serine recombinases[1,20,28,29].

For large serine integrases, differential stabilization of tetramers vs. dimers is accomplished by a subdomain (termed the coiled-coil or CC) of the recombinase itself that can mediate either activating inter-dimer interactions or inhibiting intra-dimer interactions[30–36]. In addition to the N-terminal catalytic domain, LSIs contain two DNA-binding domains, DBD1 and DBD2, also referred to as a recombinase domain (RD) and a zinc ribbon domain (ZD), respectively[4,34] (Figure 1C and 2C). The regulatory CC is inserted within DBD2 and has a hydrophobic tip and a flexible base. Whether intra- or inter-dimer CC-CC interactions are favored is determined by the details of the DNA site that each subunit is bound to, and by the presence or absence of the RDF[34].

An initial key to understanding directionality was provided by the structure of the C-terminal portion of phage LI integrase bound to half of an *attP* site, which revealed that the two DBDs recognize different sequence motifs in a modular fashion[30]. The authors deduced that the difference between *attP* and *attB* sites is that the DNA motifs for DBD2 binding are 5 bp, or half a DNA turn, closer to the center of *attB* sites than of *attP* sites[34]. Furthermore, they proposed that this positioning of DBD2 allows the CC within it to (1) promote synapsis of *attP* and *attB* sites via a "handshake" between CC's of subunits bound to different DNAs and (2) after subunit rotation, lock the dimer-bound *attL* and *attR* products via a handshake between CCs of subunits bound to the same DNAs. RDF binding was proposed to unlock the dimer-bound *att*L and *att*R complexes, allowing re-formation of an active synaptic complex[34,35]. In agreement with this idea, despite high variability in sequence and predicted structure among known RDFs, they share a common predicted binding site near the base of the CC-motif of their cognate LSIs[37–39]. However, this model for RDF function lacks a structural underpinning and struggles to explain the unidirectionality of the excisive recombination reaction (that is, it explains why the RDF activates *attL* x *attR* recombination but not why it inhibits *attP* x *attB* recombination).

To understand the structural basis of directionality control, we determined six cryo-EM structures of *B. subtilis* phage SPbeta integrase (SprA) in the presence and absence of its cognate RDF, SprB[40] and with different *att* sites, at resolutions ranging from 4.16 to 7.18Å. The resulting cryo-EM maps unveil conformational intermediates along both reaction pathways (Figure 2 and 3), and provide molecular insights into how RDF binding redirects the CC to dictate a very different synaptic complex for the excisive vs. the integrative reaction. We also found that the RDF allows CC-CC interactions to conformationally lock the *attP* – bound product dimer

(Figure 2B). Thus, our results provide direct evidence to support and expand upon prior models for directionality control in LSI/RDF systems[31,32,36,41].

Overall structures of synaptic complexes during integrative and excisive recombination

We prepared active synaptic complexes of SPbeta integrase (Int) with and without its cognate RDF, trapped in the double-strand-break intermediate state in which Ser 22 of each subunit is covalently attached to a 5' DNA end. To do so, we used synthetic *att* sites with two T:T mismatches in the central dinucleotide between the two cleavage sites (we termed these modified sites *attLmm*, *attPmm* and *attBmm*) (Figure S1, S9 and S10). These sites trapped the covalent intermediate because serine recombinases do not generally religate mismatched ends[42]. An 8bp A-tract was added to the outer ends of the P-type half-sites to help distinguish them from B-type sites (although that proved unnecessary in the final maps)[43]. To simplify the excisive complex, we used two copies of *attLmm*, based on previous data showing that the only functional difference between WT *attL* and *attR* sites is their asymmetric central dinucleotide (Figure S2)[44]. To prevent accidental loss of the small RDF protein, we fused it to the C-terminus of the integrase, with a flexible linker between the two[45]. These samples yielded structures of the pre- and post-rotation states for both the integrative and the excisive synaptic complexes, as well as two additional structures that help understand conformational trapping of recombination products.

The highest resolution structures were determined for the SPbeta int-RDF fusion protein in complex with *attLmm*. We found three distinct complexes in these samples (Figure 2A and B). Most particles contained synaptic complexes in the covalent intermediate state, which could be divided into two classes that differ by an ~180° rotation of one half of the complex relative to the other (Figure 2A). While the rotation in these samples is not constrained by any covalent bonds, there are likely energy wells at the 0° and 180° positions that are deep enough to favor these positions but shallow enough to allow rotation. Analysis of the conformational heterogeneity of the particles captured in these two representative states indicated that swiveling was not among the major components. Instead, the major contributions to heterogeneity were flexibility in the exact angle between the outer ends of the paired DNAs and partial destabilization of CC-CC interactions (see supplementary movie). The global resolution was 4.82Å for the pre-rotation state (with local resolution extending to 4.16Å) and 5.00Å for the rotated state (Figure 2A and S3). A smaller class of particles contained product *attP*-bound dimers (Figure 2B and S3). We could not identify *attB*- bound dimers, perhaps due to weaker binding to *attB* under our experimental conditions.

We also determined structures of synaptic complexes formed by mixing SPbeta Int with *attPmm* and *attBmm* in the absence of the RDF. As above, both pre- and post-rotation particle classes were found with global map resolutions of 7.18Å and 6.92Å, respectively, and local resolution extending to 4.69Å (Figure 3A and S4). Additionally, a third class of particles contained a tetramer of Int synapsing two *attP* sites yielding a 4.81Å map (Figure 3B and S4).

Overall, in all our synaptic complex structures, the two cleaved DNAs lie on the outside complex, and are held together not only by tetramerization of the catalytic domains, as seen previously for small serine recombinases[22], but also by inter-dimer handshakes between the CC motifs (Figure 4 and S5). The C-terminus of the last helix of the catalytic domain (αE) and DBD1 are in close contact with the DNA substrate near the cleavage site and positioned

similarly in all half-sites (Figure S6). The architecture of the *attP*-side complexes resembles that of the LI-*attP* half site complex structure[30]. However, when bound to *attB* half sites, the SPbeta Int DBD2 is 4 bp closer to the center (Figure S7). This displacement, as predicted[35] rotates DBD2 with its inserted CC motif to the opposite face of double helix – the critical functional difference between *attP* and *attB* half-sites. In all of the synaptic complexes, the two halves of each cleaved DNA are roughly co-linear, but the phosphoserines are ~24Å apart (measured between P atoms), ~10Å further than they would be in religated B-form DNA. This suggests that a conformational change may be needed to bring the 3' OH groups close enough for religation. The *attB* half sites tilt away from the catalytic domains, as does the DNA in most serine recombinase complexes[21,22,25,46,47], but the *attP* half sites tilt towards the catalytic domains. It is unclear if this reflects an intrinsic feature of *attP* or a "pulling" effect of the inter-dimer CC-CC interactions. However, 3D variability analysis and 3D flexible reconstruction did show some flexibility in relative orientation of the DNAs (supplemental movie).

The RDF dictates different synaptic complexes for integration and excision

Comparison of the integrative and excisive synaptic complexes shows that the CC – CC interactions occur on opposite sides of these complexes (Figures 4 and S5). Nevertheless, the local interactions between the tips of the CCs are quite similar in both cases, and similar to those reported for LI integrase, adopting a front-to-back CC-CC dimer, including a few degrees of rotational flexibility in the contact angle[32]. RDF binding to Int does not simply disrupt the "handcuffed" *attL*- and *attR*-bound dimers that are the product of integration – it stabilizes a very different trajectory of the CC and promotes the formation of a different synaptic complex for excision.

Our structures show that there are two hinges between DBD2 and the CC tip (Figure 4). The RDF binds both DBD2 and the inter-hinge segments, thus restraining the flexibility of hinge 1 but not hinge 2. (Figure 4). When the RDF is absent, both hinges adopt very different conformations. The inter-hinge segment that is N-terminal to the CC retains its helical conformation and packs against DBD2 in the same location where a loop of the RDF binds, effectively shielding a small hydrophobic region (Figure 4C and D). Part of the inter-hinge segment that is C-terminal to the CC repacks to form an extension of a helix within DBD2 itself (Figure 4A). Both the experimental maps and AlphaFold models support this helical extension in the absence of the RDF. These observations suggest that the RDF may destabilize the integration-competent trajectory of the CC as well as stabilize the excision-competent trajectory.

We employed an *E. coli*-based promoter inversion assay to investigate the functional importance of Int-RDF interactions (Figure 5A)[37,48]. Int flips a promoter flanked by *attP* and *attB* sites to drive GFP rather than RFP expression, while the Int – RDF fusion acts on the *attL* and *attR* sites in the product plasmid, flipping the promoter back to RFP expression. Use of the Int – RDF fusion avoids artifacts due to unequal protein expression or weakened protein – protein affinity. Figure 5B shows that the WT proteins drove recombination to completion after only 2 hours of expression. Deletion of the N-terminal helix of the RDF (Int-RDF ΔN-ter helix 1-13), which interacts only with DBD2, had no effect on RDF function in this assay. In contrast, the RDF's C-terminal helix (Int-RDF ΔC-ter helix 37-58), which packs between DBD2 and the inter-hinge region of the CC, is essential: when this helix was deleted, the fusion protein acted as a slow version of WT Int without RDF, eventually converting all substrates to the *attL*/att*R* configuration. Similar results were obtained for two less drastic changes in the C-terminal helix: pairwise mutations L37D/F40D and A45D/L49D. The first pair flank Int F532, which was previously shown to be important for directionality control[49], and the second pair interact directly

with the inter-hinge segment. These data support the hypothesis that restraining the inter-hinge region of the CC is critical to RDF function.

We also tested the effects of lengthening the hinges in SPbeta integrase. Hinge 1 was lengthened by inserting GGSGSSG into both the outgoing and returning chain (Figure 5B) and hinge 2 was similarly lengthened by inserting GGSGSG (see Figure 5B). The effects of elongating hinge 1 were minor: both integrative and excisive reactions were slower, and the int-RDF fusion was less cleanly switched – excisive recombination was still favored, but not as decisively. Elongating hinge 2 had no detectable effect on Int or Int-RDF function in our assays. Although elongating the CC of LI integrase by ~two helical turns weakened recombination efficiency and directionality[50], our results suggest that the most important function of the RDF is to redirect the trajectory of the CC within the synaptic complex rather than simply to restrain how far it can extend.

CC – CC interactions can lock product dimers in both reaction directions

Examination of the structures in Figure 2 and 3 shows that for both the integrative and excisive complexes, rotation to the product state moves the CCs of the product dimers closer together, which could facilitate swapping CC-CC interaction partners from inter- to intra-dimer, thus locking the product conformation. These observations not only provide structural support for the integrative recombination model[35] but also dictate an updated excisive recombination model, diagrammed in Figure 6.

In the integrative complexes, all four CCs lie on the same side of the tetramer after rotation, and the ambiguities in the map in the region for their tips suggest mixed populations of interactions. The *attL*-bound product dimer may be modeled based on the *attP* - *attP* tetramer complex found on the same grids (Figure 3B). In that complex, diagonal CC-CC interactions are formed on each side of the tetramer, facilitated by the rotation of one CC towards what would be the *att*B half-site in the real product tetramer. Combining that *attP* half with an *attB* half from the real product complexes closely juxtaposes the CC tips of the resulting *attL* site (Figure 3C). The overall architecture of these structures agrees with the prescient modeling of Rutherford and Van Duyne [35].

Although the appearance of tetramers bound to two copies of *attP* was surprising, we found that SPbeta Int – *attPmm* complexes do elute as tetramers in size exclusion chromatography (SEC) (Figure S9). The diagonal CC-CC interactions most likely prevent recombination between two *attP* sites even when they are synapsed. In contrast, SPbeta Int-RDF - *attPmm* complexes elute as *attP* bound dimers in SEC (Figure S9).

In the excisive complexes, the RDF redirects the CC such that the second hinges of both *attP*-bound subunits lie on the same "top" surface of the complex after subunit rotation (Figure 2A). Only a simple rotation about these hinges is then needed to bring the tips of those CCs together, as seen in our low-resolution maps of *attP*-bound Int-RDF dimers (Figure 2B). It is unclear whether or not *attB*-bound Int-RDF dimers can also be handcuffed by CC-CC interactions, but only the handcuffing of *attP*-bound dimers is required to explain the long-standing puzzle of how the reversal of excisive recombination is inhibited (Figure 6 and S9).

The Protein – DNA interactions are partially modular

The overall organization of the three major LSI domains (cat, DBD1, and DBD2) is not altered by the presence or absence of the RDF. We therefore used the best-resolution maps, those for

the excisive pathway synaptic complexes, to examine the details of the protein – DNA interactions (Figure 7 and S6). In comparing *attP*- vs. *attB*- half sites, density for the phosphate groups was clear enough to be certain that DBD2 lies 4bp closer to the center on the *attB* half-sites. Although analysis of many LSI binding site sequences predicts a 5bp shift[34,35,38,50], SPbeta appears to be an outlier, and the outer portions of our binding sites do align better if a 4bp shift is assumed (Figure S6 and S7). The quality of our maps prevents an extensive analysis of the structural basis for sequence specificity. Overall, direct interactions between protein side chains and DNA bases are sparse, implying that sequence recognition may rely heavily on water-mediated contacts not seen at this resolution and on indirect readout of sequence-dependent variations in the structure and flexibility of the DNA.

The local interactions between DBD2 and DNA are similar on both types of *att* site, despite its shift relative to the rest of the complex. The most prominent DNA-binding feature of DBD2 is a long beta hairpin that lies in the major groove – this hairpin is longer than in phage LI integrase[30] and its tip interacts with the bases and the phosphodiester backbone of the conserved TG at the outer edges of the binding sites. Proper placement of DBD2 is important in distinguishing *attP* from *attB*[51], and Figure S7 shows that this probably depends not only on the presence of conserved sequence motifs in the correct place, but on their absence in the wrong place.

The interdomain linker and the left side of DBD1 (as shown in Figure 7E) do not interact similarly with *attP* and *attB*. The linker binds in the minor groove and widens it, introducing a larger kink in *attB* than in *attP* (Figure 7B, C, and E). This kink facilitates an additional contact between the N-terminus of a helix on the left side of DBD1 and the phosphodiester backbone of *attB*. Even the large side chain of Y237, which DBD1 inserts into the major groove, makes different interactions with the two half sites. The different bending of the DNA is also facilitated by the narrowing of the minor groove in this region seen in *attP* but not *attB*, which correlates with an AT-rich sequence in *attP* (Figure 7A): AT-rich sequences intrinsically favor a narrowed minor groove[52]. In contrast, the right side of DBD1, where the more conserved nucleotides are found, does interact similarly with both types of binding site. Sequence-specific interactions with some of those bases may be made by the long beta-hairpin loop that extends toward the catalytic domain and inserts two R side chains into the major groove. That loop is variable among LSIs, and its major groove-interacting tip is absent in LI integrase, where DBD1 lacks significant sequence specificity[41]. This elongated loop could be a target for directed evolution to enhance or alter sequence specificity. The C-terminal portion of αE, which docks into the minor groove, also likely contributes to indirect readout of the DNA sequence near the cleavage sites, as it does in LI integrase[41].

Additional protein-protein interactions not previously predicted

The SPbeta synaptic complexes include additional protein-protein interactions that were not predicted from partial structures and previous modeling (Figure S8). The long DNA-binding loop of one subunit's DBD1 could form a salt bridge with the catalytic domain of the subunit bound to the DNA on the opposite side of the synapse, potentially helping to stabilize the synaptic complex. When the RDF is present, the CC from the *attP*-bound subunit contacts a small hydrophobic patch on DBD1 of the same subunit. When the RDF is absent, a different region of the CC from the *attP*-bound subunit lies within contact distance of the catalytic domain of the opposite subunit, again potentially stabilizing the synapse. Interestingly, that region of the catalytic domain is involved in activating contacts from other proteins or DNAs in small serine recombinases[20].

Discussion

Understanding the control of reaction directionality for the LSIs has been a long-standing challenge. Our structures of SPBeta Int and Int-RDF synaptic complexes trapped during integrative and excisive recombination provide direct evidence for an updated model that explains unidirectionality in both pathways. These structures reveal how interactions between CC subdomains can initially stabilize productive synaptic complexes and then rearrange to conformationally trap product complexes. The trajectory of the CC is determined not only by which of the two different types of *att* half-sites the modular domains of the integrase are bound to, but also by the presence vs. absence of the RDF protein. Redirection of the CC by the RDF unlocks the *attL*- and *attR*- bound products of integrative recombination, favors a synaptic complex with different geometry from that used in integrative recombination, and allows trapping of the *attP*-bound product of excisive recombination. The use of structurally different synaptic complexes for integration vs. excision is conceptually similar to the regulation of phage lambda integrase (a tyrosine recombinase) even though the two systems differ wildly in mechanism and complexity[19,53].

The work presented here also provides a framework for engineering LSI-RDF pairs and their cognate *att* sites for use as genetic switches and genome editing tools. The amino acid sequences of protein segments that adopt different interactions and/or conformations in the excisive vs. integrative pathways may reflect evolutionary compromises between pathways – not only to balance the optimal efficiency in both directions, but also to balance optimal inhibition of the "wrong" reaction in both cases. They are therefore likely to be hotspots for mutations that optimize reaction efficiency in just one direction, as seen in the clustering of efficiency mutations in CC region reported in [54] and for mutations that loosen directionality control, as seen in CC region in [36]. Our work also highlights that analysis and optimization of *att* site sequences should consider not only what is required for affinity and specificity, but also what is required to differentiate *attP* vs. *attB* half-sites, which is critical for the remarkable directionality of these molecular devices.

Acknowledgments and Funding

We thank the members of Research Computing Center (RCC) at the University of Chicago for providing their services and high performance computing cluster, Beagle3 for data storage and processing. Cryo-EM datasets were collected at the University of Chicago Advanced Electron Microscopy Core Facility (RRID:SCR_019198). Molecular graphics and analyses performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases. This publication is primarily supported by the National Science Foundation and UK Research and Innovation [collaborative grant NSF/BIO 2223480 and UKRI/BBSRC BB/X012085/1 to P.A.R. and F.J.O]. Funding for open access charge: National Science Foundation.

Author contributions

Conception and design, Methodology: PAR, FJO, HS

Funding acquisition: PAR, FJO

Data acquisition: HS, JRF, YP, TPR

Data analysis: PAR, FJO, HS, JRF

Drafting and revision: PAR, FJO, HS

Competing interests

The authors have declared no competing interest.

Data and materials availability

All EM movies have been deposited at the Electron Microscopy Public Image Archive (EMPIAR). The cryo-EM maps and coordinates of the complexes determined in this study have been deposited into the Electron Microscopy Data Bank (EMDB) and the Protein Data Bank (PDB), respectively, with the following accession codes: EMDB-47288 and 9DXH for Int in pre-rotation state; EMDB-47289 and 9DXJ for Int in post-rotation state; EMDB-47290 and 9DXK for Int *attP*-bound tetramer; EMDB-47284 and 9DXD for Int-RDF in pre-rotation state; EMDB-47286 and 9DXF for Int-RDF in post-rotation state; EMDB-47287 and 9DXG for Int-RDF *attP*-bound dimer.

Methods

Cloning and plasmid construction

The DNA sequence of SPBeta Integrase was optimized for expression in *E. coli* and cloned between the NdeI and XhoI sites of pET-28a(+) (Novagen). The proteins expressed from this vector have an N-terminal hexahistidine tag followed by a thrombin cleavage site (the tag was not cleaved off, but was not visible in the density maps). The DNA sequence encoding the SPbeta integrase–RDF fusion protein was made by joining the coding sequences for SPbeta integrase and its RDF via a 54 bp DNA linker sequence between the SpeI and XhoI sites in pET28-a(+). The DNA linker sequence is translated to a flexible 18-residue linker (TSGSGGSGGSGGSGRSGT) between the C terminal residue of SPbeta integrase and the second amino acid residue in the RDF. Mutants of Integrase and Integrase-RDF were generated using site-directed mutagenesis, and all constructs were verified by whole plasmid sequencing of the expression vectors.

Protein expression and purification

Expression plasmids containing SPBeta Int or SPBeta Int-RDF were used to transform *E. coli* BL21(DE3)pLysS competent cells. The transformed cells were grown at 37 ℃ in LB broth supplemented with 50µg/ml Kanamycin and 100 µM $ZnSO_4$, until OD_{600} reached 0.7. Protein expression was induced with isopropylβ-D-1-thiogalactopyranoside (IPTG) at final concentration of 0.5 mM, and the cultures were grown overnight at 20 ℃. The cells were pelleted by centrifugation at 8,000 r.p.m. in a Fiberlite F9-6X1000LEX rotor for 7 minutes at 4 ℃ and stored at -20C.

Cell pellets were resuspended in 50ml per liter of original culture of Ni A buffer (50mM Sodium Phosphate, 1M NaCl, 5% Glycerol, 1mM Tris(2-carboxylethyl) phosphine (TCEP), pH 7.5) plus 1 tablet of complete Mini EDTA-free protease inhibitor cocktail (Roche) and 200 µg/ml lysozyme, then lysed by ultrasonication in 3 1-minute pulses in an ice bath. The lysate was clarified by centrifugation at 18,000 r.p.m. in an SS-34 rotor for 1 hour at 4 ℃. The supernatant was collected, filtered and then loaded on a 5ml Ni Sepharose HisTrap HP affinity column (Cytiva). The column was washed with 95% Ni A / 5% Ni B buffer (Ni B = Ni A plus 500 mM

Imidazole pH 7.5) until the baseline was flat, then eluted by a gradient from 5% to 100% Ni B) over 40' with 2ml/min and 2ml/fraction. The protein peak fractions were analyzed on a Novex 8-16% Tris-Glycine Minigel (Invitrogen). The cleanest fractions were pooled and dialyzed into Ni A buffer at 4C O/N, then rechromatographed on the Ni column as before. The best fractions were pooled, then diluted into Heparin A buffer (20mM MES, 5% Glycerol, 1mM TCEP, pH 5.5) to a final salt concentration of 200-300mM, and then loaded onto a 20ml Heparin column (Heparin FF, 16/10, Cytiva), washed with 15% Heparin B buffer (Heparin A buffer plus 2M NaCl, pH5.5) and eluted with a gradient from 15% to 75% B over 90' at 2ml/min and 2ml/fraction. The cleanest peak fractions were and concentrated to a small volume, then dialyzed into 20 mM Tris-Cl pH 8, 200 mM NaCl, 2 mM TCEP, and 20% glycerol. The protein concentration was determined by A_{280} and the presence of contaminating endonucleases was ruled out by incubating the sample with supercoiled pUC19 and 10mM $MgCl_2$ followed by agarose gel electrophoresis.

List of DNA substrates

Names	Sequences (5'→3')
*attLsym*_top	CATGTCATTAATATCAGTACAGATATAGCTGTATATTAAGATACTTACT ACATATCTAAAAAAAAAA
*attLsym*_bot	TTTTTTTTTTAGATATGTAGTAAGTATCTTAATATACAGCTATATCTGTA CTGATATTAATGACATG
*attRsym*_top	AAAAAAAAAAAGAAAAGTAGTAAGTATCTTAAAAAACAGATATAGC TGTATCTCCTGTGAACACAAT
*attRsym*_bot	ATTGTGTTCACAGGAGATACAGCTATATCTGTTTTTTAAGATCATTAC TACTTTTCTTTTTTTTTTT
*attLmm*_top	CATGTCATTAATATCAGTACAGATTTAGCTGTATATTAAGATACTTACT ACAAAAAAAA
*attLmm*_bot	TTTTTTTTGTAGTAAGTATCTTAATATACAGCTTTATCTGTACTGATATT AATGACATG
*attPmm*_top	AAAAAAAAGGTAGTAAGTATCTTAAAAAACAGATTTAGCTGTATATT AAGATACTTACTAC
*attPmm*_bot	GTAGTAAGTATCTTAATATACAGCTTTATCTGTTTTTTAAGATACTTAC TACTTTTTTTTT
*attBmm*_top	CATGTCATTAATATCAGTACAGATTTAGCTGTATCTCCTGTGAACACA ATAAAAAAAA
*attBmm*_bot	TTTTTTTTATTGTGTTCACAGGAGATACAGCTTTATCTGTACTGATATT AATGACATG

For the fluorescently labeled oligos, some of the oligos listed above were custom ordered with 5' fluorescein label or 3' Cy5 label from Integrated DNA technologies (IDT). Complementary oligos were mixed at a final concentration of 1 mM in TE buffer containing 100 mM NaCl and annealed by incubating at 80 °C for 20 minutes. Afterward, the water bath was turned off and samples were left in it to cool to room temperature.

In vivo integrase mediated inversion assay

The design and protocol for the *in vivo* recombination assay were similar to those previously described[37,48].The substrate for *in vivo* intramolecular recombination (inversion) is a reporter plasmid (kanamycin resistance) in which *attP*/*attB* or *attR*/*attL* sites are in inverted orientation and flank a constitutive promoter that drives either expression of GFP or RFP (Figure 5a). To assay recombination activities *in vivo*, *E. coli* DS941 cells were cotransformed with a substrate plasmid and an arabinose-inducible expression vector, pBAD33 (chloramphenicol resistance) containing the gene of interest (Int, Int-RDF fusion or its mutants cloned between NdeI and XhoI sites). To do so, 2µl of each plasmid at 20ng/µl was incubated with 100ul chemically competent DS941 cells on ice for 5', then heat shocked at 37C for 5', then iced for 2'. 200µl LB+0.2% glucose was then added, cells were allowed to recover 37C for 2 hours, then added to 10ml LB+0.2% glucose+50µg/ml kanamycin (kan) +30µg/ml chloramphenicol (cam) and grown overnight at 37C. All liquid cultures for these experiments were shaken at 200 rpm. Glucose was added to suppress protein expression, and we noted that the overall experiment worked best if the overnight culture reached an OD_{600} of > 1.5 – if not it was best to restart with freshly made competent cells.

200µl of the overnight culture was used to inoculate 10ml LB + 0.2% Glucose + kan (50 µg/ml) + cam (30µg/ml), then grown at 37C for 1.5 hours, at which point OD_{600} was ~0.4-0.5. Cells were spun down to remove the glucose, then resuspended in 10ml LB + 0.2% L-Arabinose + kan (50 µg/ml) + cam (30 µg/ml) and incubated at 37C for 2 hours or overnight, after which the plasmids were recovered by miniprep. 1µl of miniprepped plasmid was then used to transform 50ul chemically competent DS941 cells: 5' on ice, 5' heat shock at 37C and 2' on ice then recovery in 500µl LB at 37C for 90'. We choose to isolate and then retransform the product plasmids to separate out individual copies of the substrate plasmid, which carries a pSC101 origin and can be present in multiple copies per cell. Two different volumes of each culture (20µl and 100µl) were then plated LB + Kan (50 µg/ml) agar plates and incubated at 37C overnight. Plates were visualized in a chemi-doc (Bio-Rad, Hercules, CA, USA) with blue epi illumination and a 530nm±28nm bandpass filter for GFP visualization and red epi illumination with a 695nm±55nm bandpass filter for RFP. Exposure times were set at 0.025s and 0.25s respectively.

Recombination and binding assays

For binding and recombination assays, protein and DNA were used in ratio 2:1 (final concentrations: DNA, 0.2 µM; protein, 0.4 µM). The reactions were performed in a buffer containing 20 mM Tris-HCl pH 8.0, 100 mM NaCl, 10 % glycerol, 50 µg/ml BSA, 0.5 mM TCEP. Samples were incubated at 37 ℃ for 1 hr.

After the incubation, the samples were divided into two sets. To assess recombination, 10 % of Sodium Dodecyl Sulfate (SDS) and 6X DNA loading dye (6X Gel loading dye, purple—New England Biolabs) were added, resulting in final concentrations of 1.1% SDS and 1X loading dye. Both sets of samples were separated on Novex 6% TBE gels (Invitrogen) in 1X TBE running buffer at 160V for 45 minutes or until the dye front reached near the bottom of the gel. The gels were visualized using a Chemidoc gel imager (Bio-Rad).

Sample preparation for single particle cryo-EM

To obtain the complexes, purified and concentrated SPBeta integrase (Int) or Int-RDF fusion proteins in storage buffer (20 mM Tris-Cl pH 8, 200 mM NaCl, 2 mM TCEP, and 20 % glycerol) were mixed with annealed mismatched dsDNA. Specifically, Int was mixed with *attLmm* and Int-RDF was mixed with *attPmm* and *attBmm* at final protein to DNA concentrations of 80 µM

to 40 µM, respectively, in low salt buffer (20 mM Tris-Cl pH 8, 100mM NaCl). The samples were incubated at 37 ℃ for 1 hour and purified using size-exclusion chromatography (SEC) on a Superose 6 10/300 GL column, Cytiva pre-equilibrated with low salt buffer (20mM Tris-Cl pH 8, 100mM NaCl).

The selected fractions containing the complexes were concentrated to 5mg/ml using using Amicon Ultra Centrifugal Filter Concentrator (30kDa, MilliporeSigma), and then adjusted to 2mg/ml with a low salt buffer (20mM Tric-Cl pH 8, 100mM NaCl). Immediately prior to blotting and vitrification, fluorinated octyl maltoside (FOM, Anatrace) and TCEP were added to final concentration of 0.02 % (w/v) and 0.125 mM, respectively. 3.5 µl of the sample was applied to a freshly glow discharged CFlat 4.0/1.0 300-mesh grid, blotted for 5.5 seconds with blot force set to 0, and plunged into liquid ethane using a Vitrobot™ MK IV (FEI) at 18 ℃ and 100 % relative humidity.

Data collection for single particle cryo-EM

Imaging of particles was performed using a Titan Krios G3i (Thermo Fisher Scientific) operating at 300 kV and equipped with a Gatan K3 direct detection camera. Data collection was conducted at the Advanced Electron Microscopy Facility, the University of Chicago. Each movie consisted of 54 frames with a total exposure of approximately 65 electrons per Å^2. The defocus range was set at -1.0 µm to -2.5 µm. Additional acquisition and imaging parameters are detailed in Extended Data Table 1.

Image processing

All data processing steps were conducted using cryoSPARC (v4.6.0)[55] and are summarized in Figures S3 and S4. For both datasets, micrograph movies were aligned, summed, and dose-weighted; contrast transfer function (CTF) parameters were estimated using patch based CTF estimation from cryoSPARC. Micrographs with significant non-vitreous ice contamination or outliers in motion correction and CTF estimation were excluded.

Initial particle picking was performed using the blob particle picker to generate two-dimensional (2D) classifications. After a second round of 2D classification, selected classes showing intact complexes were used as templates for automated particle picking. False-positive or poorly defined particles were filtered out after additional rounds of 2D classification. The remaining particles were used to train a Topaz[56] model, which were applied to the combined micrographs for refined particle picking.

Particles classes showing clear secondary structures were iteratively classified, and three to five three-dimensional (3D) ab initio models were generated with C1 or C2 symmetry. Classes representing intact synaptic complexes were selected and further refined through heterogenous and non-uniform refinement steps. Resolutions for both Int and Int-RDF datasets were improved through focused refinement steps, where particles from both pre- and post-rotated classes were combined, resulting in the alignment of the half complex. Reference-based motion correction was applied, followed by particle subtraction to remove the unaligned part of the other half complex, and the final set of particles were used for local refinement.

To refine flexible regions of the complexes, built-in 3D variability analysis[57] and 3D flex reconstruction[58] from cryoSPARC[55] were performed to address disordered regions and deconvolute the heterogeneous conformations of the complex into top three components ranked by significance in conformational variability. Using the outputs of these two analytical programs, we generated movies (see supplementary movies) depicting the various conformational variabilities in the consensus maps and in the focused maps within UCSF Chimera (v.1.17.3)[59]. Four maps from the last stage of the non-uniform refinement cycle, corresponding to pre- and

post-rotated states in both the Int and Int-RDF datasets, were used as representative consensus maps. Two focused refined half-complex maps were then combined using the corresponding consensus map as a guide to generate composite maps for these rotational states. The final maps were sharpened using cryoSPARC's built-in map sharpening tools, and the composite maps for pre- and post- rotated states were further processed with deepEMhancer (v.0.14)[60].

Model building, refinement and validation

Alphafold2-multimer[61,62] was used to create the starting model of the synaptic complex. Specifically, the complex was predicted in overlapping fragments to allow flexibility in relative domain arrangements. These fragments were subsequently positioned and aligned according to the cryo-EM map using ChimeraX (v.1.4)[63]. Overlapping regions were manually corrected, adjusted, and refined in COOT (v.0.9.8.8)[64]. All of the DNA could be modeled except for the 3' most of the two overhanging Ts created by the DNA cleavage reaction. While most protein residues were built, some regions containing flexible loops, the 20aa long N-terminal tag, 28aa C-terminal residues and the linker connecting the Int and RDF were not built because they were disordered in the maps. The register of the main chain was carefully checked and corrected based on bulky residues.

B-form DNA starting models for the DNA substrates were built in ChimeraX and fitted into the cryo-EM map with local restraints using COOT[64]. ISOLDE (v.1.4)[65], embedded in UCSF ChimeraX[63], was used to introduce phosphoserine linkages between the protein and DNA, and to correct most of peptide bond geometry, rotamer, and Ramachandran outliers.

Manual model building and refinement were performed iteratively using COOT[64], ISOLDE[65], and PHENIX (v.1.21)[66] real-space refinement. Both unsharpened and sharpened maps, as well as 3D-flex reconstructed maps, were used in conjunction with previously solved truncated structures and Alphafold2[61,62] predicted models to guide the model building process. Geometry restraints for zinc-binding sites were applied based on a zinc-binding model to ensure accurate representation of these regions.

Molecular graphics and structural analyses were conducted using UCSF ChimeraX[63] and PyMOL (v. 3.1.3)[67].

References

1. Smith, M. C. M. Phage-encoded Serine Integrases and Other Large Serine Recombinases. *Microbiol. Spectr.* **3**, 1128/microbiolspec.mdna3-0059–2014 (2015).
2. Olorunniji, F. J., Rosser, S. J. & Stark, W. M. Site-specific recombinases: molecular machines for the Genetic Revolution. *Biochem. J.* **473**, 673–683 (2016).
3. Stark, W. M. Making serine integrases work for us. *Curr. Opin. Microbiol.* (2017).
4. Smith, M. C. M. & Thorpe, H. M. Diversity in the serine recombinases. *Mol. Microbiol.* **44**, 299–307 (2002).
5. Khaleel, T., Younger, E., McEwan, A. R., Varghese, A. S. & Smith, M. C. M. A phage protein that binds φC31 integrase to switch its directionality. *Mol. Microbiol.* **80**, 1450–1463 (2011).
6. Ghosh, P., Wasil, L. R. & Hatfull, G. F. Control of Phage Bxb1 Excision by a Novel Recombination Directionality Factor. *PLoS Biol.* **4**, (2006).
7. Smith, M. C. M., Brown, W. R. A., McEwan, A. R. & Rowley, P. A. Site-specific recombination by φC31 integrase and other large serine recombinases. *Biochem. Soc. Trans.* **38**, 388–394 (2010).

8. Bibb, L. A. & Hatfull, G. F. Integration and excision of the Mycobacterium tuberculosis prophage-like element, φRv1. *Mol. Microbiol.* **45**, 1515–1526 (2002).
9. Bibb, L. A., Hancox, M. I. & Hatfull, G. F. Integration and excision by the large serine recombinase φRv1 integrase. *Mol. Microbiol.* **55**, 1896–1910 (2005).
10. Breüner, A., Brøndsted, L. & Hammer, K. Novel Organization of Genes Involved in Prophage Excision Identified in the Temperate Lactococcal Bacteriophage TP901-1. *J. Bacteriol.* **181**, 7291–7297 (1999).
11. Blanch-Asensio, A. STRAIGHT-IN enables high-throughput targeting of large DNA payloads in human pluripotent stem cells. *Cell Rep. Methods* (2022) doi:10.1016/j.crmeth.2022.100300.
12. Gomide, M. S. *et al.* Genetic switches designed for eukaryotic cells and controlled by serine integrases. *Commun. Biol.* **3**, 255 (2020).
13. Merrick, C. A., Zhao, J. & Rosser, S. J. Serine Integrases: Advancing Synthetic Biology. *ACS Synth. Biol.* **7**, 299–310 (2018).
14. Guiziou, S., Maranas, C. J., Chu, J. C. & Nemhauser, J. L. An integrase toolbox to record gene-expression during plant development. *Nat. Commun.* **14**, 1844 (2023).
15. Anzalone, A. V. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* **38**, (2020).
16. Durrant, M. G. *et al.* Systematic discovery of recombinases for efficient integration of large DNA sequences into the human genome. *Nat. Biotechnol.* **41**, 488–499 (2023).
17. Yarnall, M. T. N. *et al.* Drag-and-drop genome insertion of large sequences without double-strand DNA cleavage using CRISPR-directed integrases. *Nat. Biotechnol.* **41**, 500–512 (2023).
18. Vo, P. L. H. *et al.* CRISPR RNA-guided integrases for high-efficiency, multiplexed bacterial genome engineering. *Nat. Biotechnol.* **39**, 480–489 (2021).
19. Grindley, N. D. F., Whiteson, K. L. & Rice, P. A. Mechanisms of Site-Specific Recombination. *Annu. Rev. Biochem.* **75**, 567–605 (2006).
20. Rice, P. A. Serine Resolvases. *Microbiol. Spectr.* **3**, (2015).
21. Kamtekar, S. *et al.* Implications of structures of synaptic tetramers of gamma delta resolvase for the mechanism of recombination. *Proc. Natl. Acad. Sci.* **103**, 10642–10647 (2006).
22. Li, W. *et al.* Structure of a Synaptic gd Resolvase Tetramer Covalently Linked to Two Cleaved DNAs. *Science* **309**, (2005).
23. Dhar, G., Sanders, E. R. & Johnson, R. C. Architecture of the Hin Synaptic Complex during Recombination: The Recombinase Subunits Translocate with the DNA Strands. *Cell* **119**, 33–45 (2004).
24. Stark, W. M., Sherratt, D. J. & Soocock, R. Site-Specific Recombination by Tn3 Resolvase: Topological Changes in the Forward and Reverse Reactions. *Cell* **58**, 779–790 (1989).
25. Yang, W. & Steitz, T. A. Crystal Structure of the Site-Specific Recombinase Resolvase Complexed with a 34 bp Cleavage Site. *Cell* **82**, 193–207 (1995).
26. Mouw, K. W. *et al.* Sin Resolvase Catalytic Activity and Oligomerization State are Tightly Coupled. *J. Mol. Biol.* **404**, 16–33 (2010).
27. Arnold, P. H., Blake, D. G., Grindley, N. D. F., Boocock, M. R. & Stark, W. M. Mutants of Tn3 resolvase which do not require accessory binding sites for recombination activity. *Eur. Mol. Biol. Organ.* **18**, 1407–1414 (1999).
28. Johnson, R. C. Site-specific DNA Inversion by Serine Recombinases. *Microbiol. Spectr.* **3**, 3.1.13 (2015).

29. McLean, M. M., Chang, Y., Dhar, G., Heiss, J. K. & Johnson, R. C. Multiple interfaces between a serine recombinase and an enhancer control site-specific DNA inversion. *eLife* **2**, e01211 (2013).
30. Rutherford, K., Yuan, P., Perry, K., Sharp, R. & Van Duyne, G. D. Attachment site recognition and regulation of directionality by the serine integrases. *Nucleic Acids Res.* **41**, 8341–8356 (2013).
31. Fogg, P. C. M. *et al.* Recombination directionality factor gp3 binds φC31 integrase via the zinc domain, potentially affecting the trajectory of the coiled-coil motif. *Nucleic Acids Res.* **46**, 1308–1320 (2017).
32. Gupta, K., Sharp, R., Yuan, J. B., Li, H. & Van Duyne, G. D. Coiled-coil interactions mediate serine integrase directionality. *Nucleic Acids Res.* **45**, 7339–7353 (2017).
33. McEwan, A. R., Rowley, P. A. & Smith, M. C. M. DNA binding and synapsis by the large C-terminal domain of φC31 integrase. *Nucleic Acids Res.* **37**, 4764–4773 (2009).
34. Van Duyne, G. D. & Rutherford, K. Large serine recombinase domain structure and attachment site binding. *Crit. Rev. Biochem. Mol. Biol.* **48**, 476–491 (2013).
35. Rutherford, K. & Van Duyne, G. D. The ins and outs of serine integrase site-specific recombination. *Curr. Opin. Struct. Biol.* **24**, 125–131 (2014).
36. Rowley, P. A., Smith, M. C. A., Younger, E. & Smith, M. C. M. A motif in the C-terminal domain of ϕC31 integrase controls the directionality of recombination. *Nucleic Acids Res.* **36**, 3879–3891 (2008).
37. MacDonald, A. I. *et al.* Variable orthogonality of serine integrase interactions within the φC31 family. *Sci. Rep.* (2024).
38. Shin, H. *et al.* Identification of cognate recombination directionality factors for large serine recombinases by virtual pulldown. Preprint at https://doi.org/10.1101/2024.06.11.598349 (2024).
39. Alsaleh, A. *et al.* Large serine integrases utilise scavenged phage proteins as directionality cofactors. Preprint at https://doi.org/10.1101/2024.08.21.608926 (2024).
40. Abe, K., Takamatsu, T. & Sato, T. Mechanism of bacterial gene rearrangement: SprA-catalyzed precise DNA recombination and its directionality control by SprB ensure the gene rearrangement and stable expression of spsM during sporulation in Bacillus subtilis. *Nucleic Acids Res.* **45**, 6669–6683 (2017).
41. Li, H., Sharp, R., Rutherford, K., Gupta, K. & Van Duyne, G. D. Serine Integrase attP Binding and Specificity. *J. Mol. Biol.* **430**, 4401–4418 (2018).
42. Stark, W. M., Grindley, N. D. F., Hatfull, G. F. & Boocock, M. R. Resolvase-catalysed reactions between res sites differing in the central dinucleotide of subsite I. *Eur. Mol. Biol. Organ.* **10**, 3541–3548 (1991).
43. Nelson, H. C. M., Finch, J. T., Luisi, B. F. & Klug, A. The structure of an oligo(dA)·oligo(dT) tract and its biological implications. (1987).
44. Ghosh, P., Bibb, L. A. & Hatfull, G. F. Two-step site selection for serine-integrase-mediated excision: DNA-directed integrase conformation and central dinucleotide proofreading. *Proc. Natl. Acad. Sci.* **105**, 3238–3243 (2008).
45. Olorunniji, F. J. *et al.* Control of serine integrase recombination directionality by fusion with the directionality factor. *Nucleic Acids Res.* **45**, 8635–8645 (2017).
46. Mouw, K. W. *et al.* Architecture of a Serine Recombinase-DNA Regulatory Complex. *Mol. Cell* (2008) doi:10.1016/j.molcel.2008.02.023.
47. Montano, S. P. *et al.* Structural basis for topological regulation of Tn3 resolvase. *Nucleic Acids Res.* **51**, 1001–1018 (2022).

48. Olorunniji, F. J. *et al.* Control of ϕC31 integrase-mediated site-specific recombination by protein trans-splicing. *Nucleic Acids Res.* **47**, 11452–11460 (2019).
49. Abe, K., Takahashi, T. & Sato, T. Extreme C-terminal element of SprA serine integrase is a potential component of the "molecular toggle switch" which controls the recombination and its directionality. *Mol. Microbiol.* **115**, 1110–1121 (2021).
50. Mandali, S. & Johnson, R. C. Control of the Serine Integrase Reaction: Roles of the Coiled-Coil and Helix E Regions in DNA Site Synapsis and Recombination. *J. Bacteriol.* **203** (2021).
51. Singh, S., Ghosh, P. & Hatfull, G. F. Attachment Site Selection and Identity in Bxb1 Serine Integrase-Mediated Site-Specific Recombination. *PLoS Genet.* **9**, e1003490 (2013).
52. Rohs, R. *et al.* The role of DNA shape in protein–DNA recognition. *Nature* **461**, 1248–1253 (2009).
53. Van Duyne, G. D. & Landy, A. Bacteriophage lambda site-specific recombination. *Mol. Microbiol.* **121**, 895–911 (2024).
54. Fanton, A. *et al.* Site-specific DNA insertion into the human genome with engineered recombinases. Preprint at https://doi.org/10.1101/2024.11.01.621560 (2024).
55. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
56. Bepler, T. *et al.* Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. *Nat. Methods* **16**, 1153–1160 (2019).
57. Punjani, A. & Fleet, D. J. 3D Variability Analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM images. *J. Struct. Biol.* **213**, (2021).
58. Punjani, A. & Fleet, D. J. 3DFlex: determining structure and motion of flexible proteins from cryo-EM. *Nat. Methods* **20**, 860–870 (2023).
59. Pettersen, E. F. *et al.* UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
60. Sanchez-Garcia, R. *et al.* DeepEMhancer: a deep learning solution for cryo-EM volume post-processing. *Commun. Biol.* **4**, 874 (2021).
61. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
62. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
63. Meng, E. C. *et al.* Tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).
64. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
65. Croll, T. I. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr. Sect. Struct. Biol.* **74**, 519–530 (2018).
66. Afonine, P. V. *et al.* Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. Sect. Struct. Biol.* **74**, 531–544 (2018).
67. The PyMOL Molecular Graphics System, Version 3.1.3. Schrödinger, LLC.
68. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: A Sequence Logo Generator. *Genome Research* **14**, 1188-1190 (2004). https://pubmed.ncbi.nlm.nih.gov/15173120/

Extended Data Table 1 Cryo-EM Data Collection, Refinement and Validation Statistics

	SPbeta Int pre-rotation state (EMDB-47288) (PDB 9DXH)	**SPbeta Int post-rotation state (EMDB-47289) (PDB 9DXJ)**	**SPbeta Int *att*Pmm tetramer (EMDB-47290) (PDB 9DXK)**	**SPbeta Int RDF pre-rotation state (EMDB-47284) (PDB 9DXD)**	**SPbeta Int RDF post-rotation state (EMDB-47286) (PDB 9DXF)**	**SPbeta Int RDF *att*Pmm dimer (EMDB-47287) (PDB 9DXG)**
Data collection and processing						
Magnification	81,000	81,000	81,000	81,000	81,000	81,000
Voltage (kV)	300	300	300	300	300	300
Electron exposure ($e^-/Å^2$)	65	65	65	65	65	65
Defocus range (μm)	-1.0 to -2.5	-1.0 to -2.5	-1.0 to -2.5	-1.0 to -2.5	-1.0 to -2.5	-1.0 to -2.5
Pixel size (Å)	1.065	1.065	1.065	1.065	1.065	1.065
Symmetry imposed	C1	C1	C1	C1	C1	C1
Initial particle images (no)	1,364,105	1,364,105	1,364,105	2,114,654	2,114,654	2,114,654
Final particle images (no)	13,917	27,865	92,812	550,169	248,916	54,766
Map resolution (Å)	7.18	6.92	4.81	4.82	5.00	6.61
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143
Refinement						
Model resolution (Å)	7.34	7.44	7.37	4.23	4.22	12.70
FSC threshold	0.5	0.5	0.5	0.5	0.5	0.5
Map sharpening *B* factor ($Å^2$)	-368.8	-438.8	-229.6	-321.1	-395.5	-505.3
Model Composition						
Non-hydrogen atoms	42438	42440	43689	46399	46360	23758
Protein residues	2134	2134	2140	2366	2364	1182
Nucleotide	232	232	268	232	232	134
Ligands	ZN: 4	ZN: 4	ZN: 4	ZN: 4	ZN: 4	ZN: 2
B factors ($Å^2$)						
Protein	208.26	212.19	596.91	72.31	72.31	1333.16
Nucleotide	235.45	237.11	598.59	68.69	68.69	71.64
R.m.s. deviations						
Bond lengths (Å)	0.006	0.005	0.006	0.005	0.005	0.004
Bond angles (°)	0.724	0.684	0.757	0.633	0.633	0.757
Validation						
MolProbity score	2.50	2.38	2.39	1.86	1.90	2.36
Clashscore	28.11	22.46	25.88	7.96	8.63	22.42
Poor rotamers (%)	0.00	0.00	0.00	0.00	0.00	0.00
Ramachandran plot						
Favored (%)	89.46	90.50	91.93	93.39	93.23	91.23
Allowed (%)	10.21	9.17	7.93	6.41	6.69	8.77
Disallowed (%)	0.33	0.33	0.14	0.00	0.09	0.00

Main figures

Supplementary figures

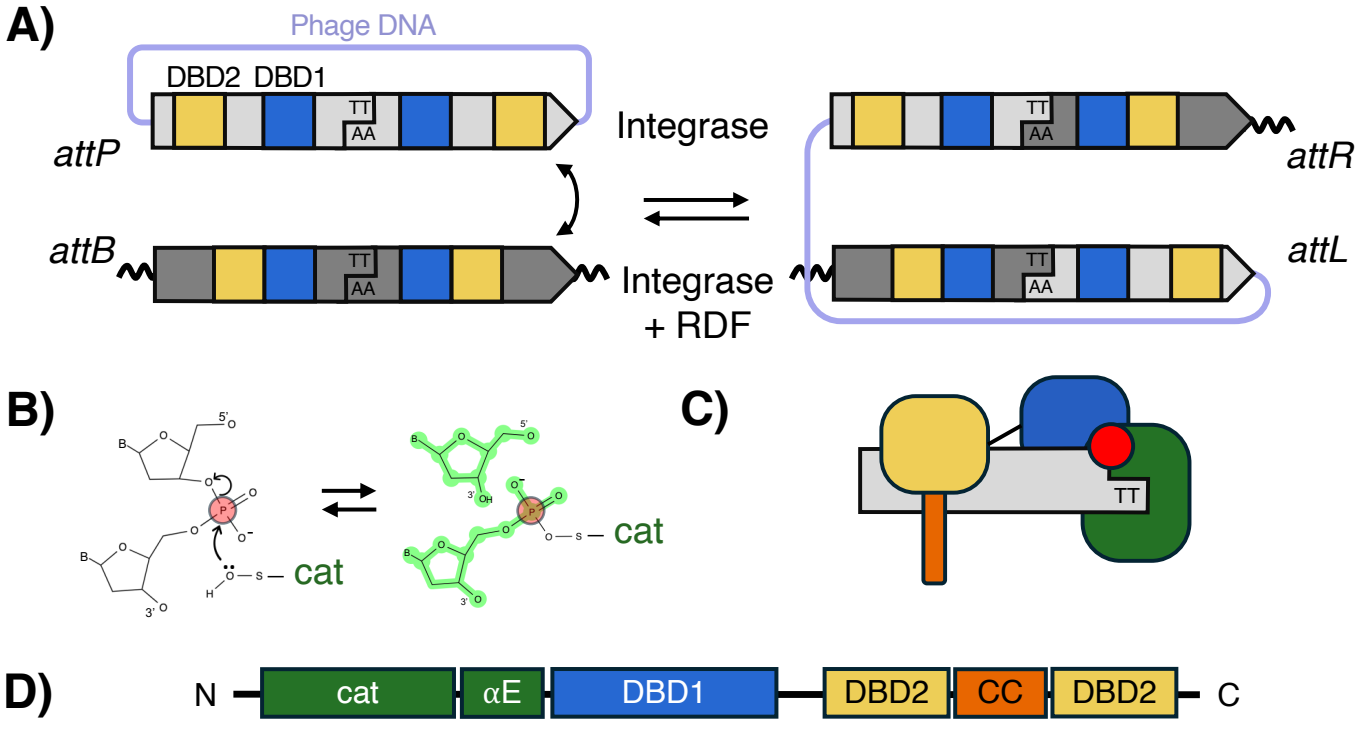


Figure 1. Large serine integrase – mediated DNA recombination.

A. Schematic diagram of the reaction. The integrase recombines a specific site within the phage (*attP*; light gray) with one found in the host chromosome (*attB*; dark gray). DNA cleavage in the center of the site creates a 2 nt 3' overhang. After exchange of the right-hand halves of the sites (as shown), the prophage is inserted within the host chromosome, flanked by hybrid sites termed *attL* and *attR*. Motifs recognized by the first DNA binding domain of the Integrase (DBD1; also known as a recombinase domain or RD) are shown in blue and motifs recognized by DBD2 (also known as zinc-binding domain or ZD) are shown in yellow. Note that *attP* and *attB* sites differ in the spacing of these two motifs. The RDF protein triggers the integrase to catalyze the reverse reaction.

B. The catalytic domain ("cat"; green) is covalently linked to the 5' ends of the cleaved DNA intermediate through a phosphoserine linkage. The phosphorous atom is indicated by a red circle here and throughout the figures.

C. Schematic representation of a large serine integrase bound along an *attP* half site and (D) domain architecture of large serine integrase which consists of catalytic domain (green), DBD1 (blue), DBD2 (yellow); and a coiled coil motif (CC; orange) that is inserted within DBD2.

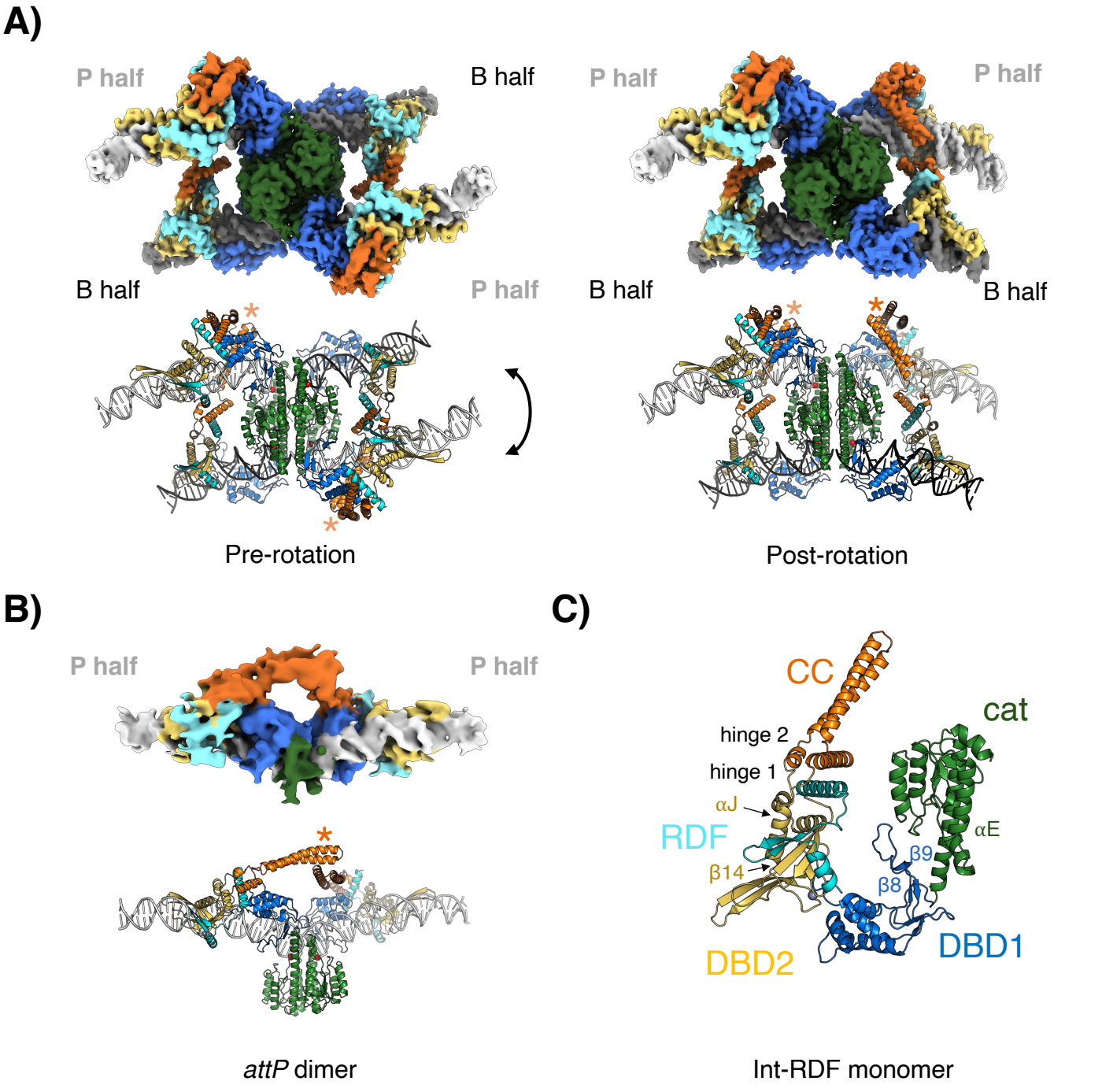


Figure 2. Structures of the Int-RDF synaptic complex during the excision (*attR* x *attL*) pathway.

A. Electron microscopy (EM) maps (top) and refined atomic structures (bottom) of the SPbeta integrase RDF fusion (Int-RDF) synaptic complexes captured in the covalent intermediate stage of excisive recombination. The pre-rotation state is shown on the left and the post-rotation state is shown on the right. Structures are colored as in Figure 1C with the RDF in turquoise. The orange asterisks indicate where the CC tips from the subunits bound to the upper and lower DNAs meet (see also the closeup in Figure 4(B)).

B. *attP*-bound product dimer structure, with orange asterisk highlighting where the CC tips of these two subunits interact to stabilize the dimer.

C. Ribbon diagram of one SPbeta Int-RDF fusion monomer isolated from the synaptic complex.

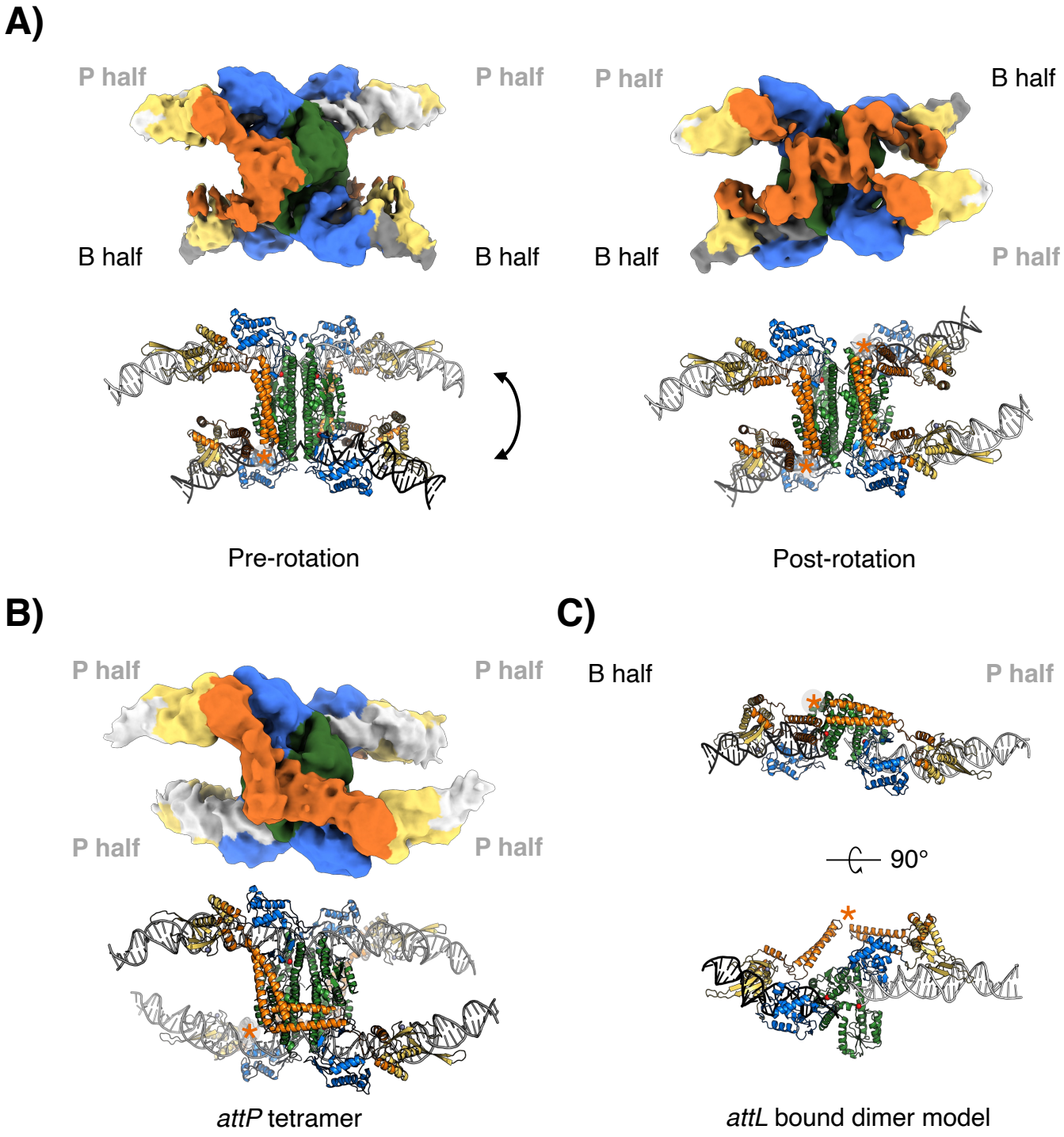


Figure 3. Structures of the synaptic complex during the integration (*attP* x *attB*) pathway (in the absence of the RDF).

A. EM maps (top) and structures (bottom) of the SPbeta integrase (Int) synaptic complexes in the pre-rotation (left) and post-rotation (right) states, captured in the covalent intermediate stage of integrative recombination. Colors are as in Figure 1C and 2, and orange asterisks mark approximately where CC:CC interactions occur. See also the closeup in Figure 4A.

B. *attP*-bound tetramer, shown in a conformationally locked state with diagonal CC:CC interactions.

C. A hybrid model illustrating a dissociated *attL*- or *attR*-bound dimer after strand exchange. This model was constructed by combining the bottom right subunit of the *attP*-bound tetramer (as shown above) and the bottom left subunit of the post-rotated product complex, which were docked together by aligning their catalytic and DBD2 domains with those of the *attP*-bound dimer structure.

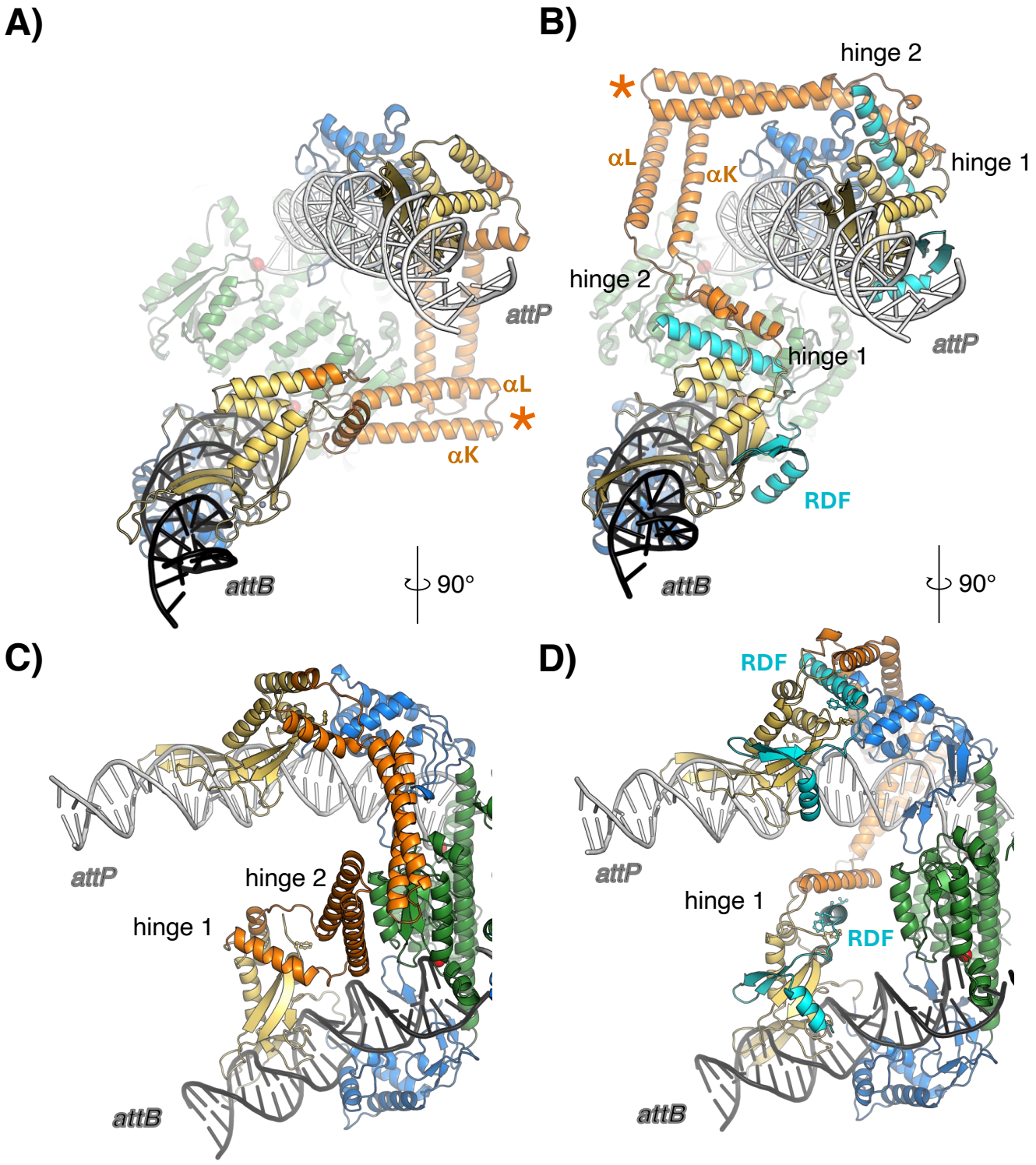


Figure 4. The RDF redirects the CC subdomain, dictating a different synaptic complex.

A. DNA-end view of the product synaptic complex in the absence of the RDF, colored as in previous figures. The orange asterisk marks where two CC motifs meet.

B. A similar view of the substrate synaptic complex in the presence of the RDF.

C. The same model as in Figure 4A, rotated ~90° about the vertical axis. Int F532 is shown as sticks (also shown in Figure 5).

D. The same model as in Figure 4B, rotated ~90° about the vertical axis. F532 and the RDF side chains that were mutated are shown as sticks (also shown in Figure 5).

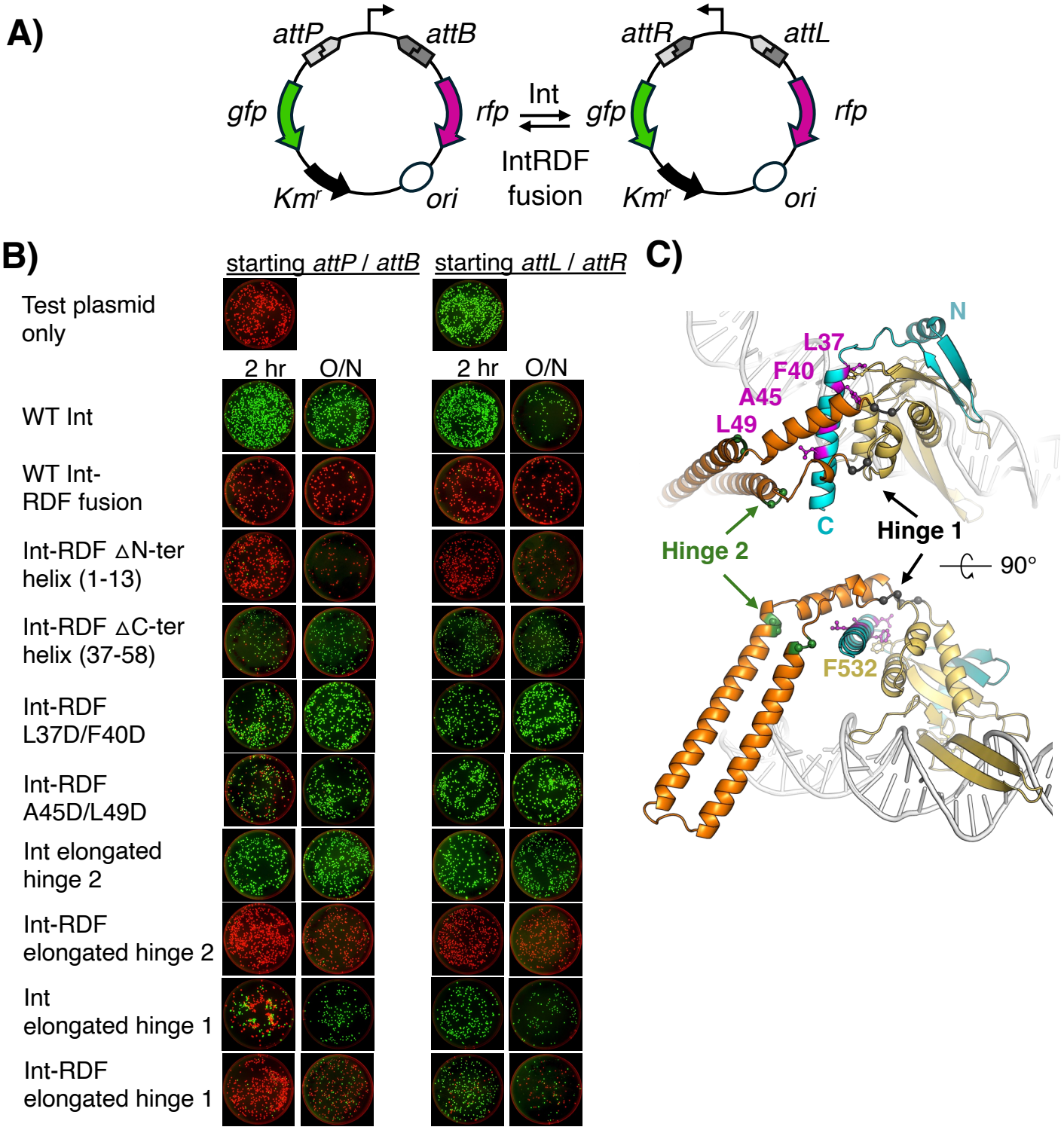


Figure 5. Mutational analysis of RDF and integrase hinge regions

A. Schematic diagram illustrating the *in vivo* DNA inversion recombination assay.

B. False-color images of plates of *E. coli* harboring the product plasmids from inversion assays. Int or Int-RDF fusions were expressed for 2 hours or overnight (see methods for details). ΔN-ter: deletion of the first helix of the RDF (Δ residues 1-13). ΔC-ter helix: RDF Δ residues 37-58. Elongated hinge 1: addition of GGSGSSG between D379 & L380 and between N479 & N480; Elongated hinge 2 : addition of GGSGSG between S401 & N402 and between D467 & T468.

C. Close-up views of DBD2 and the RDF from at *attP*-bound subunit. Side chains that were mutated are shown as sticks and the Cα atoms of residues flanking the hinge-elongating insertions are shown as spheres (gray; hinge 1, green; hinge 2).

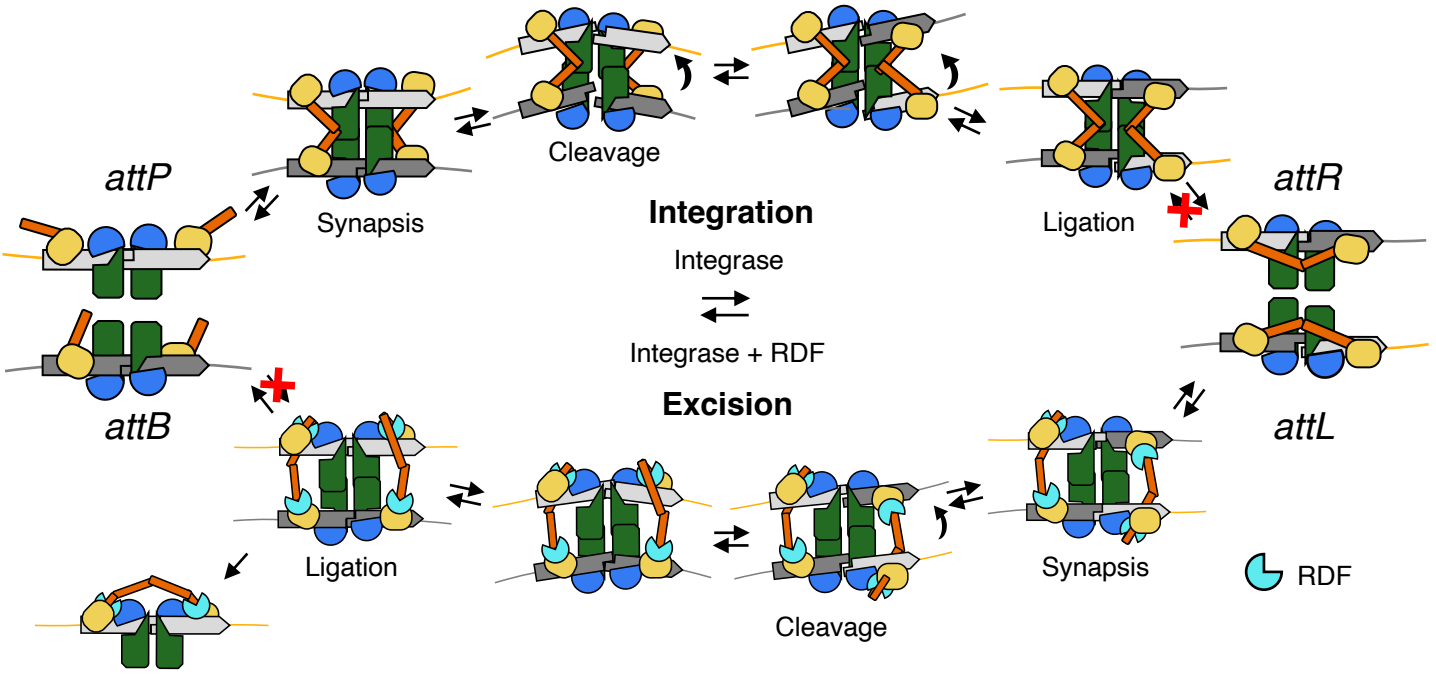


Figure 6. Schematic model of how CC-CC interactions and the RDF mediate directionality control.

Pathways for integrative (upper) and excisive (lower) recombination are diagrammed with the protein domains and DNAs colored as in previous figures. CC subdomains (orange) buttress allowed synaptic complexes by mediating interactions between the upper and lower DNAs, but they can also re-arrange after recombination to form intra-dimer interactions that trap the products and prevent the reverse reaction from occurring (red X's). The integration pathway is conceptually similar to that proposed by Rutherford and Van Duyne[35] but the CC motifs in particular are redrawn to match the experimental data presented here. The RDF redirects the trajectories of the CC subdomains so that a different type of synaptic complex (between *attL* and *attR*) and a different product dimer (*attP*-bound) are stabilized.

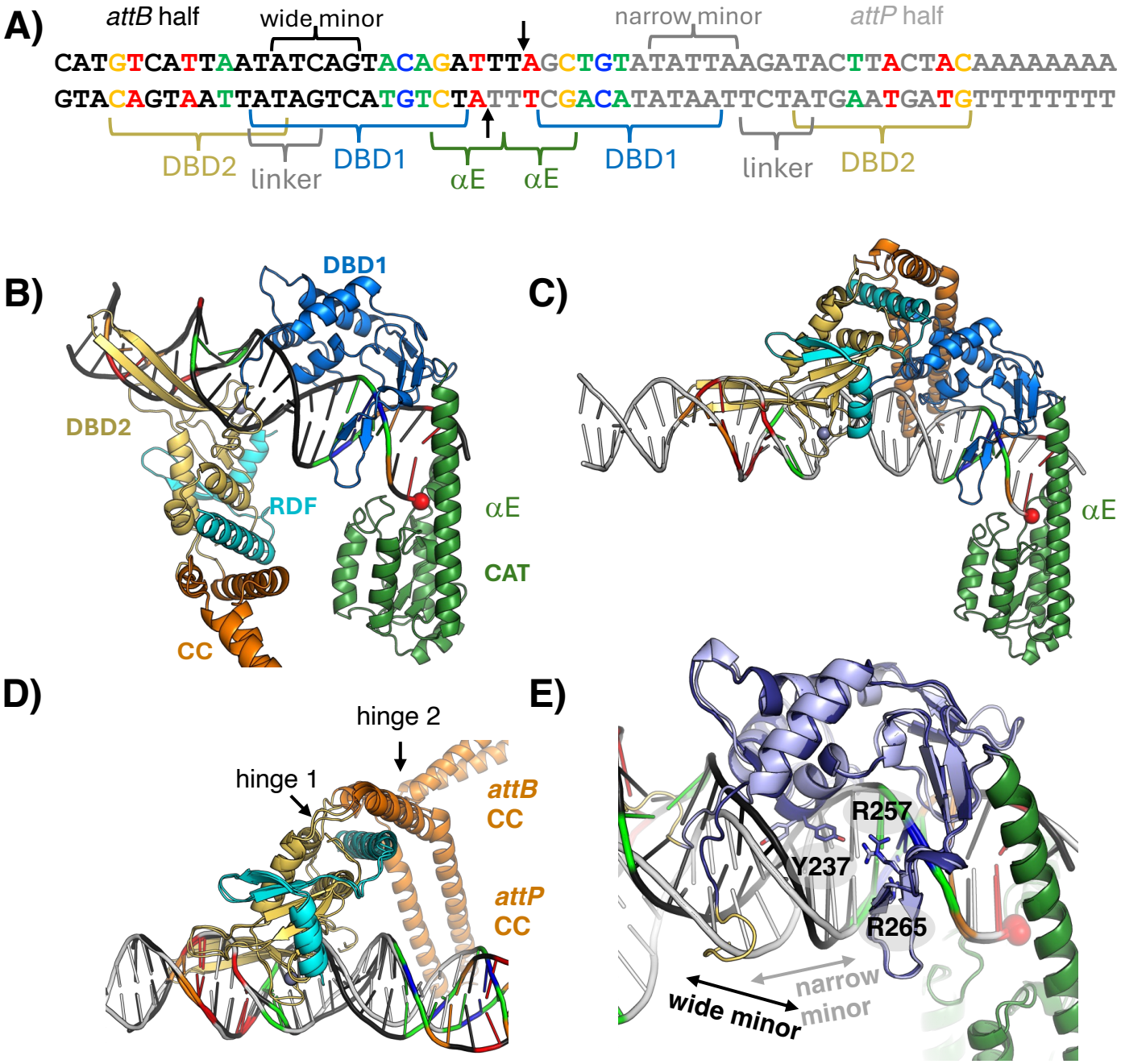


Figure 7. Interactions of DBD1 and DBD2 with the *attB* and *attP* halves of *attLmm*

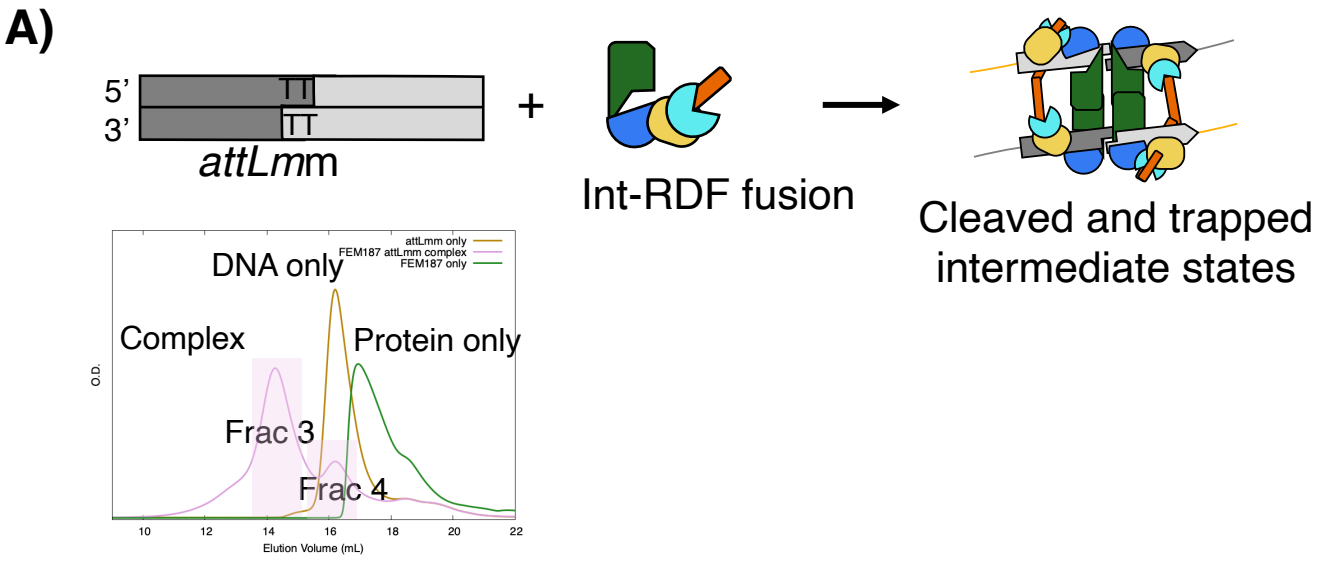
A. Specific positioning of the C-terminal domains along the *attB* and *attP* half-sites. The sequence used for structure determination is shown, including a poly-A tract added to the right end. Bases that are identical in all 4 WT half-sites are colored according to the sequence logo in Supplementary Figure 7 (reading 5' to 3'): G; orange, T; red, A; green, C; blue. Other bases in the *attB* half are black and those in the *attP* half, gray. Arrows indicate cleavage sites.

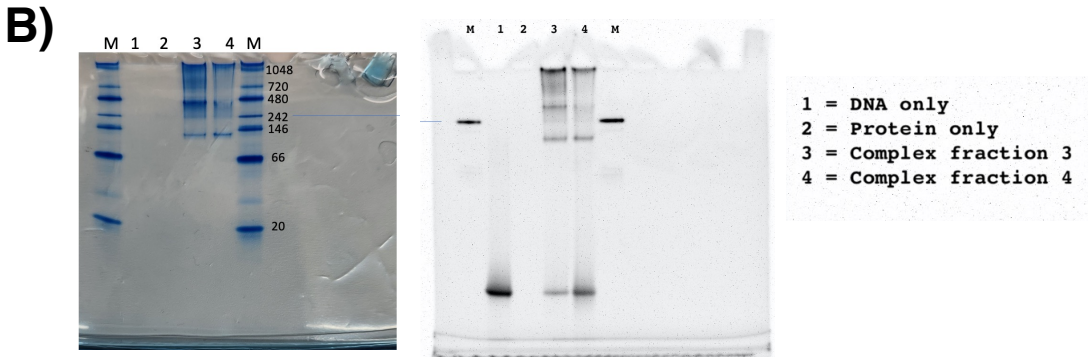
B. SPbeta Int-RDF fusion bound to the *attB* half site. DNA is colored as in (A).

C. SPbeta Int-RDF fusion bound to the *attP* half site, with the catalytic and DBD1 domains oriented similarly to those in (B).

D. Superimposed structures of *attB*- and *attP*- bound DBD2 domains, aligned based on DBD2, excluding the CC motif.

E. Superimposed structures of *attB*- and *attP*- bound DBD1 domains, aligned based on DBD1. The *attB*- bound DBD1 is in a darker shade of blue.





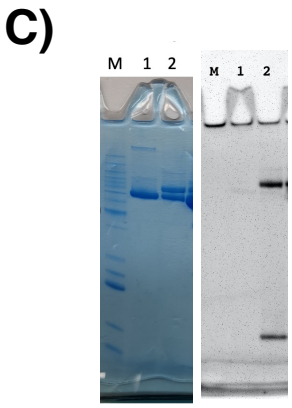
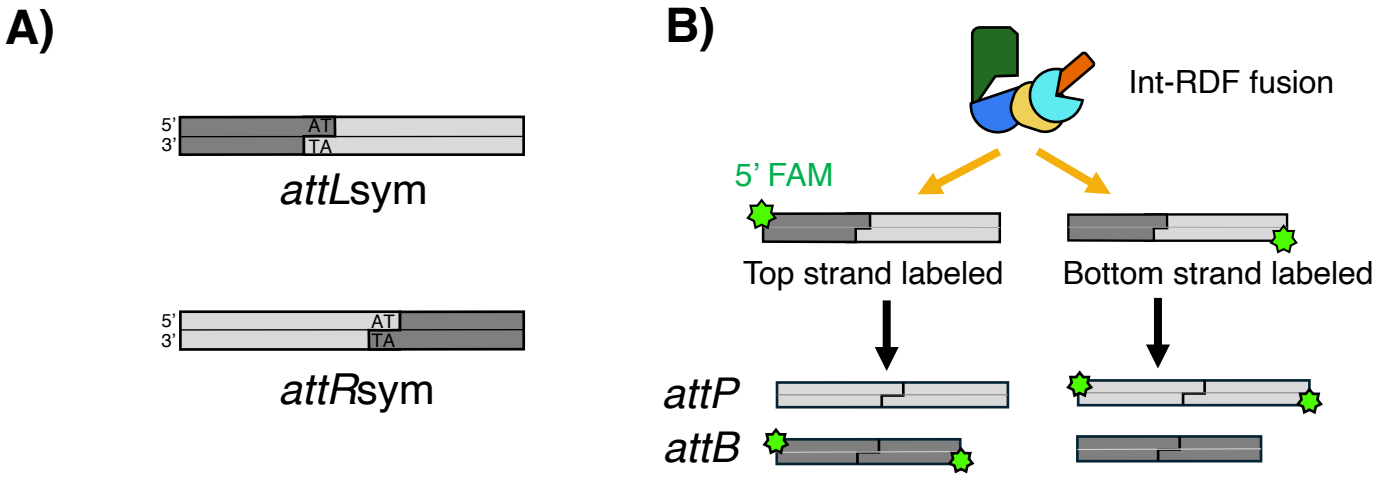


Figure S1. Mismatches in the central dinucleotide trap the covalent intermediate

A. Elution profiles of *attL*mm-bound SPbeta Int-RDF complex (in pink), *attL*mm only (in brown), and SPbeta Int-RDF only (in green) from size-exclusion chromatograph (SEC).

B. Analysis of covalent intermediate products on the native PAGE gel. Gel Code blue stained (left) and ethidium bromide stained (right). The positively charged protein only enters the gel when covalently attached to the DNA (fraction 3 and 4 are in lane 3 and 4).

C. The covalent intermediate product yield analyzed on an SDS-PAGE gel. Gel Code blue stained (left) and ethidium bromide stained (right). The covalent intermediate product corresponds to the upper band of the duplet protein band (comparing lane 1 vs. lane 2).



C)

Lanes	1	2	3	4	5	6	7	8	9
Int +RDF	-	-	+	-	+	-	+	-	+
DNA	*attB* 50bp	*attL* 67bp		*attL* 67bp		*attR* 67bp		*attR* 67bp	
Label location	Top	Top		Bottom		Bottom		Top	

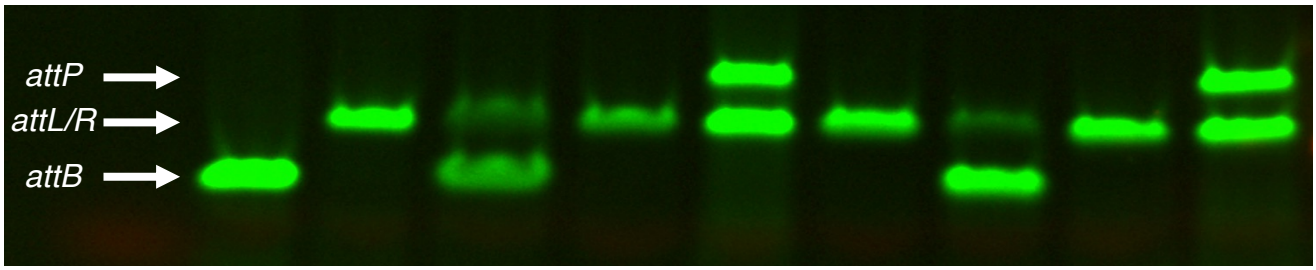


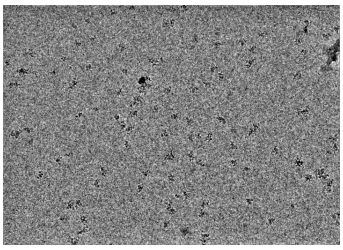
Figure S2. The SPbeta Int-RDF fusion protein recombines *attL*sym and *attR*sym substrates with comparable efficiency.

A. The *attL*sym and *attR*sym substrates (67 bp) carried the WT *attL* and *attR* sequences except that an extension was added to the *attP* side and the central dinucleotide (naturally TT) was symmetrized (now AT) to allow productive recombination between two copies of *attL* or two copies of *attR*.

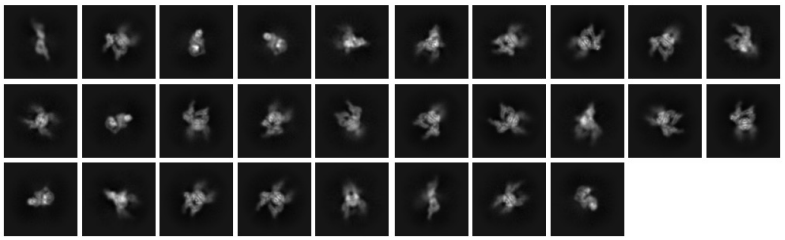
B. For each DNA substrate, the 5' end of either the top or bottom strand is labeled with fluorescein (marked with a green star). The Int-RDF fusion protein mediates the recombination of *attL*sym × *attL*sym or *attR*sym × *attR*sym substrates to generate rearranged *attB*s (50 bp) and *attP*s (84 bp) products.

C. Fluorescently labeled recombinant products are visualized on the gel, enabling the detection of one of the two recombination products.

A)



B)



16,646 movies

Motion correction
Patch CTF estimation

12,618 micrographs

Topaz train
Topaz extract = 320 pix, 1.065Å/pix

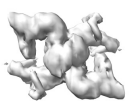
2,114,654 particles

2D classification (2 rounds)
Ab-initio reconstruction (C1 or C2)

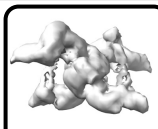
Template generation

Blob picker
Extract = 160 pix, 2.13Å/pix
2D classification (2 rounds)
Template picker = 280Å
Extract = 160 pix, 2.13Å/pix
2D classification (2 rounds)
Ab-initio reconstruction

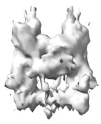
Heterogeneous refinement (C2)



299,233 particles

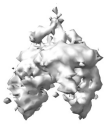


250,936 particles



109,815 particles

166,789 particles

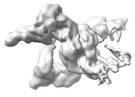


98,213 particles

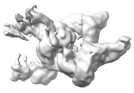
Non-uniform refinement (C1)

54,766 particles

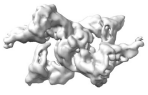
3D classifications



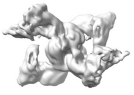
87,089 particles



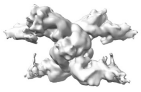
85,355 particles



63,532 particles

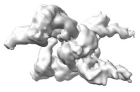


63,257 particles

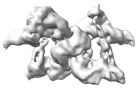


69,744 particles

68,464 particles



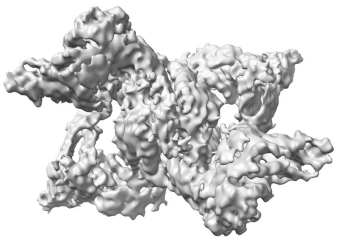
67,015 particles



45,713 particles

Non-uniform refinement (C1)

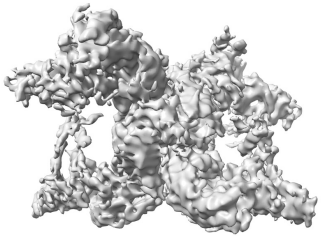
attP
attB



attB
attP

Pre-rotation

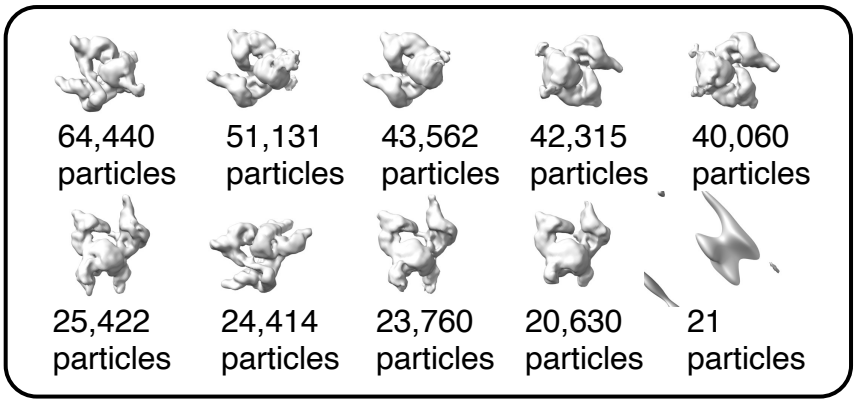
attP
attB



attP
attB

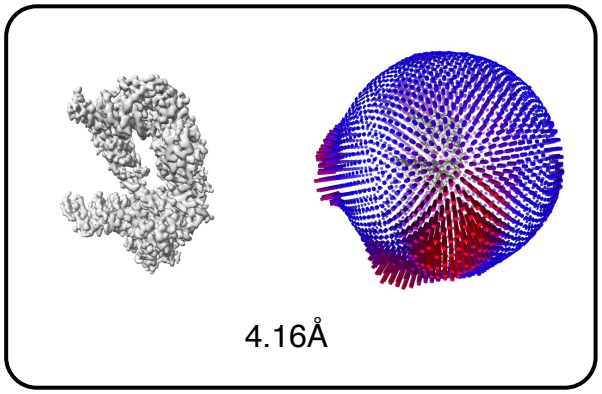
Consensus maps

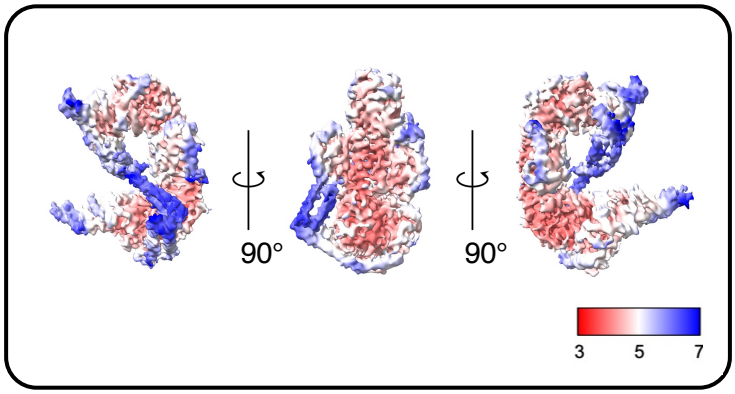
Post-rotation



Pre- and post-rotation particles are combined to generate a half complex

Homogeneous refinement
Particle subtraction
Reference based motion correction
Local refinement





Flex data preparation
Custom Flex mesh preparation
Flex train 8 dimensions
Flex reconstruction 40 iterations

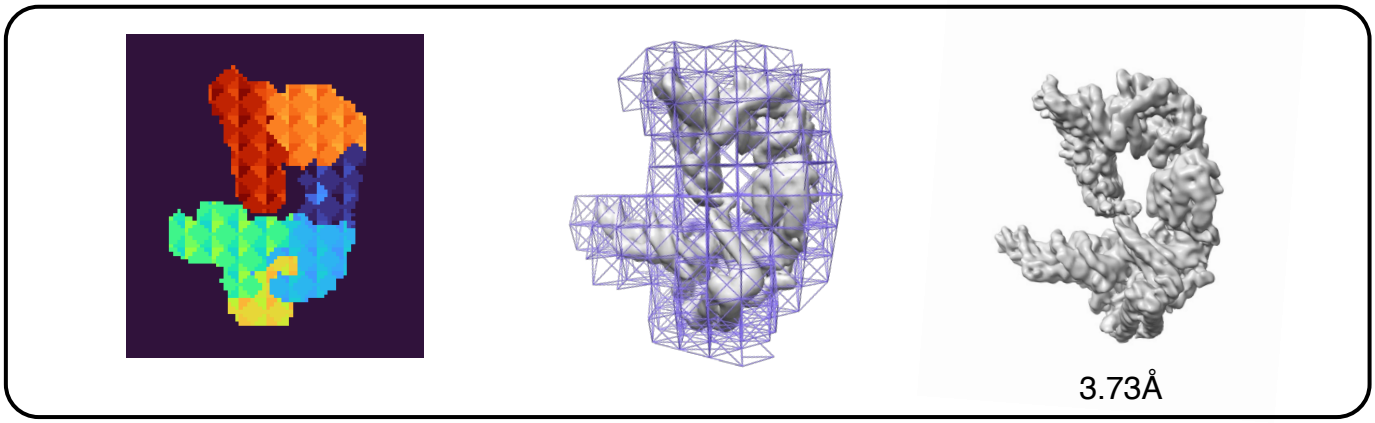
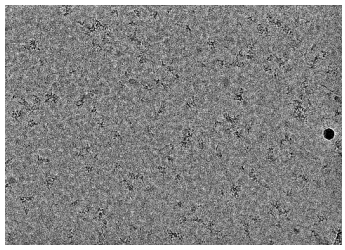


Figure S3. Data processing flow chart for the SPbeta Int-RDF data.

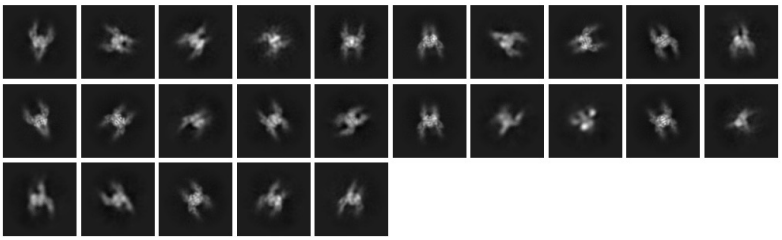
A. A visual overview of the steps involved in the processing of Int-RDF data, starting from raw movies to the final 3D maps. The final maps include three consensus maps, one focused map, and two composite maps. Two rotational states (pre- and post-rotation) are indicated with gray and black labels, respectively; the *attP* bound dimer map is shown in navy colored label. The consensus maps are outlined in light blue.

B. 2D classes illustrating various views of the synaptic complexes and their heterogeneity.

A)



B)



8,904 movies

Motion correction
Patch CTF estimation

7,893 micrographs

Topaz train
Topaz extract = 320 pix, 1.065Å/pix

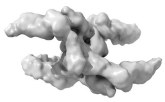
214,125 particles

2D classification (2 rounds)
Ab-initio reconstruction (C1)

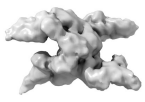
Template generation

Blob picker
Extract = 160 pix, 2.13Å/pix
2D classification (2 rounds)
Template picker = 280Å
Extract = 160 pix, 2.13Å/pix
2D classification (2 rounds)
Ab-initio reconstruction

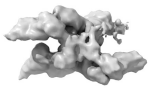
Heterogeneous refinement (C1)



103,146 particles

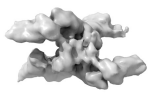


60,381 particles

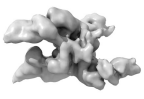


50,598 particles

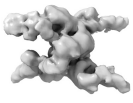
3D classifications



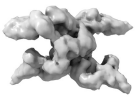
26,831 particles



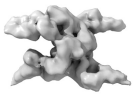
23,267 particles



21,429 particles

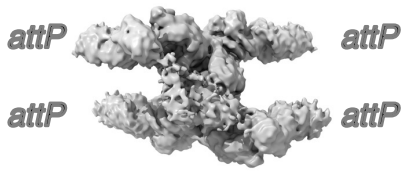


19,832 particles

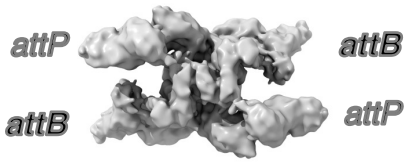


19,620 particles

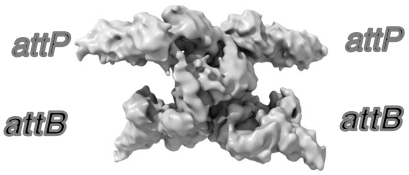
Non-uniform refinement



92,812 particles



27,865 particles



13,917 particles

Consensus maps

Focused refinements

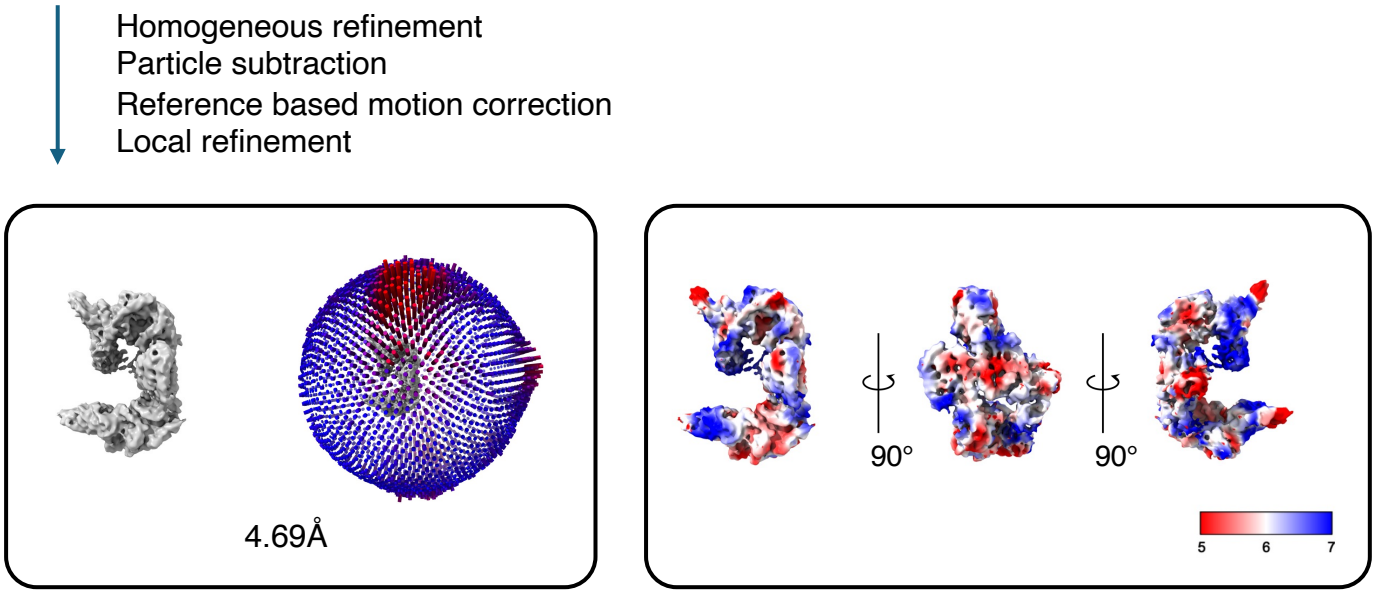


Figure S4. Data processing flow chart for the SPbeta Int data.

A. A visual overview of the steps involved in the processing of Int data, starting from raw movies to the final 3D maps. The final maps include three consensus maps, one focused map, and two composite maps. The two rotational states (pre- and post-rotation) are indicated in gray and black labels, respectively; the *attP* bound tetramer map is shown in a navy label. The consensus maps are outlined in light blue.

B. 2D classes illustrating various views of the synaptic complexes and their heterogeneity.

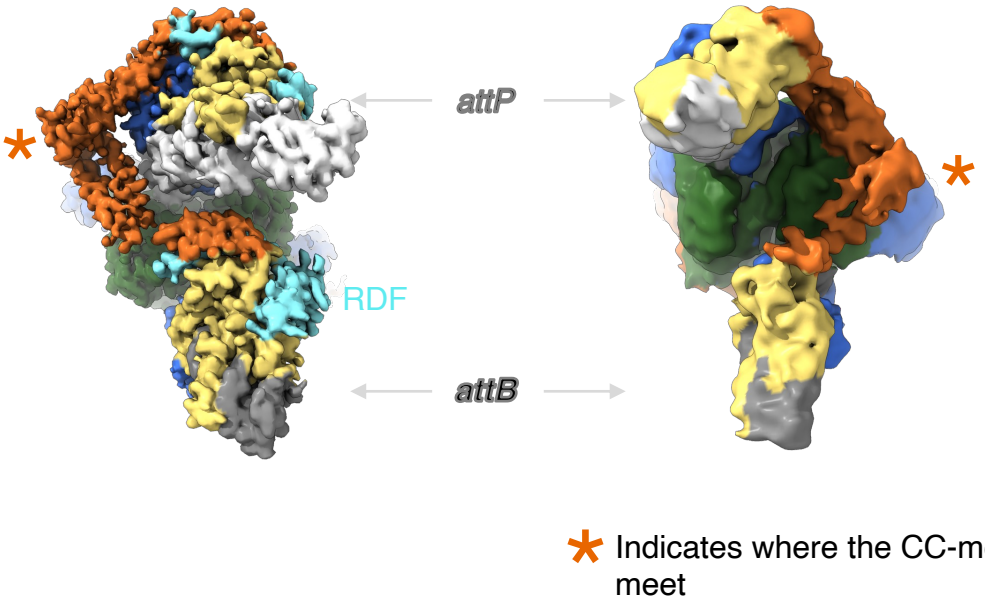


Figure S5. Redirection of the CC-motif trajectory by RDF confers distinct synaptic complexes. Side-by-side comparison of EM maps in the absence (right) and presence (left) of RDF (in cyan). In the presence of RDF binding, the CC-motifs are oriented in the opposite side of the synaptic complex. The location where the two CC-motifs meet is indicated by an orange asterisk.

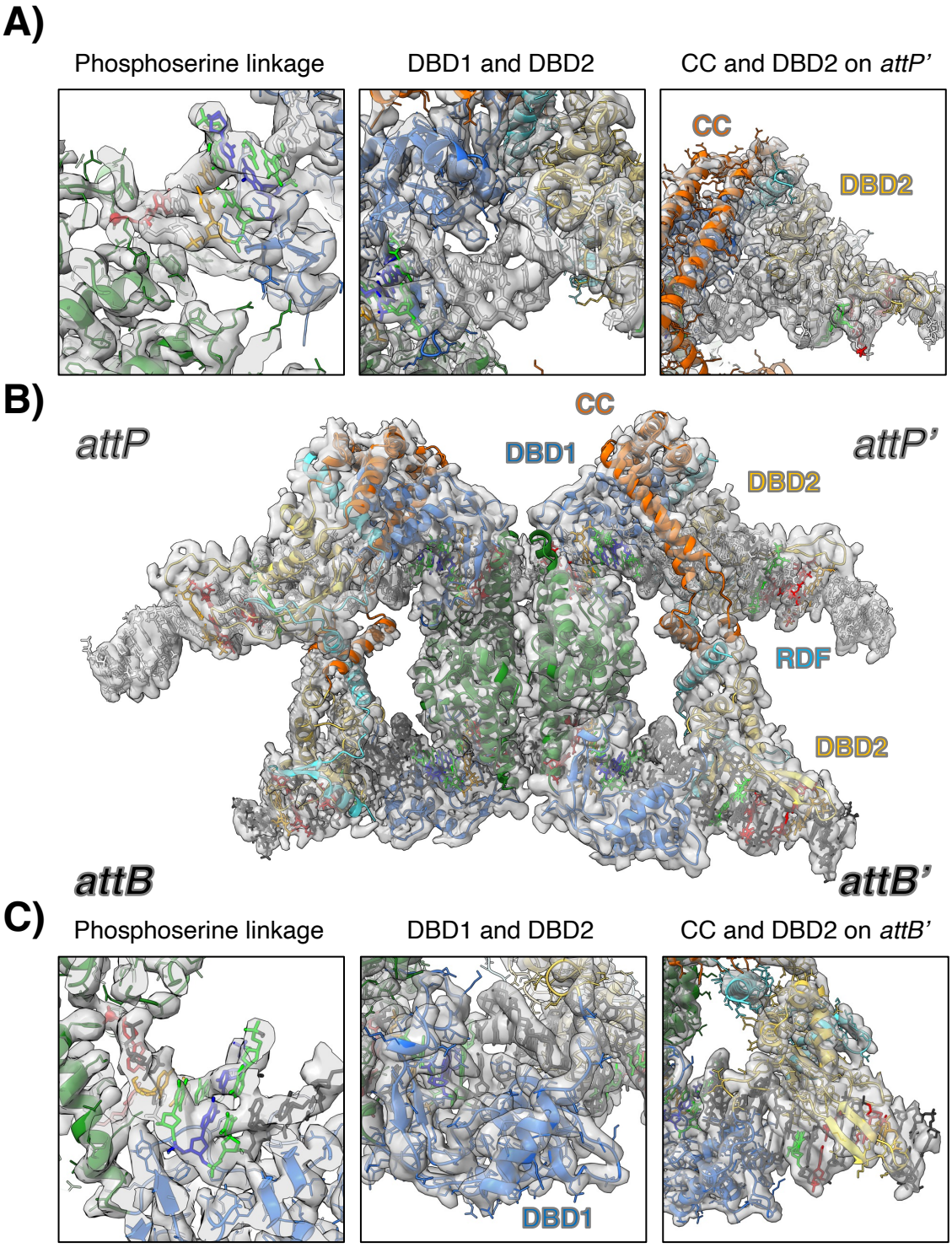


Figure S6. Experimental maps supporting protein—DNA interactions

A. Colors are the same as in Figures 1B and 7 (for the DNA). The Phosphoserine linkage (red sphere) of S22 to the 5' end of *attP*' (left), C-terminal end αE and N-terminal region of DBD1 bound to *attP*' near the cleavage site; attachment of DBD1 and DBD2 domains on *attP*' (middle); attachment of DBD2 domain near the outer end of the *attP'* half site.

B. Overall structure of the Int-RDF synaptic complex in the post rotation state.

C. Phosphoserine linkage of S22 to the 5' end of *attB*' (left), attachment of DBD1 and DBD2 domains on *attB*' (middle); attachment of the DBD2 domain near the outer end of the *attB*' half site.

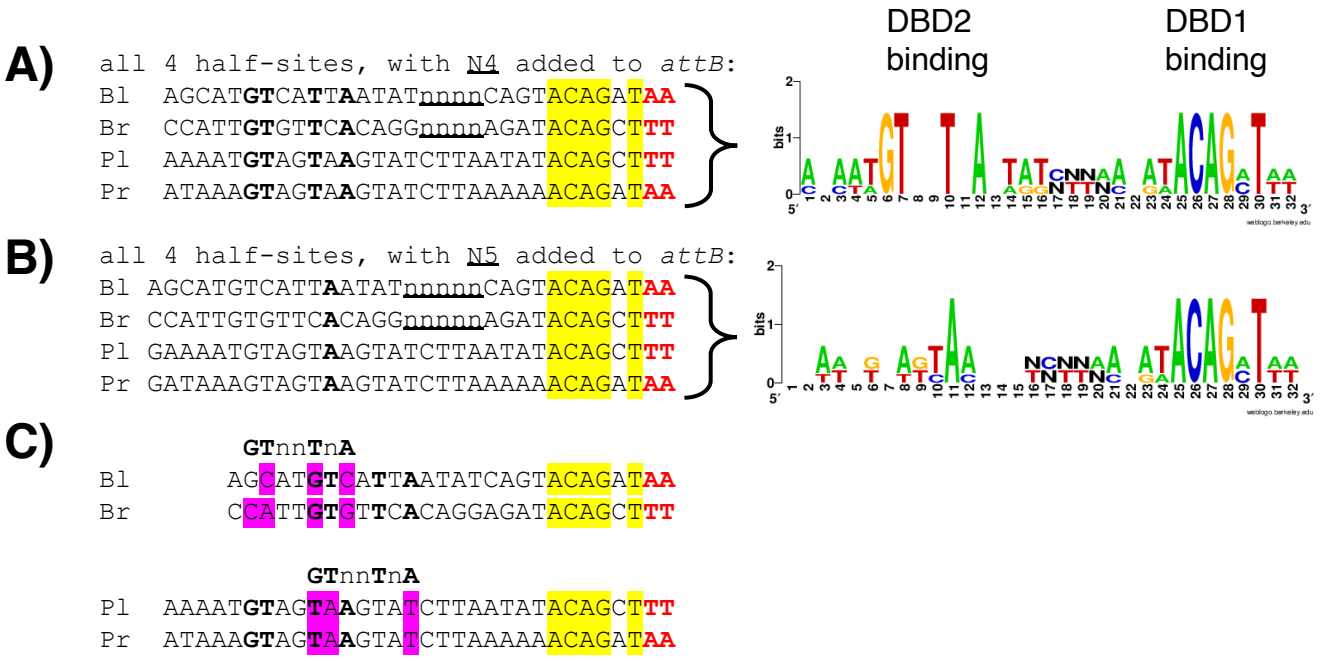


Figure S7. Differentiation of the two types of *att* site.

A. Alignment of all 4 natural half-sites, assuming a 4-bp offset for the DBD2-specific motif between B- and P- type half sites, as supported by the experimental maps. The *att* sites for phage SPbeta were taken from the genome sequence of *B. subtilis* (NCBI Accession: NZ_CP051860.2). Bl and Pl are the two halves that combine to form *attL*; likewise Br and Pr comprise *attR*. The central dinucleotide is highlighted in red. Nucleotides that are identical in all 4 half-sites in the DBD1-binding region are highlighted in yellow, while those in the DBD2-binding region are shown in bold. A sequence logo (https://weblogo.berkeley.edu/logo.cgi)[68] for these alignments is shown on the right.

B. A similar alignment of SPbeta half-sites, but with N5 added between motifs, as may be more common among other large serine integrases[34,38].

C. *attP* and *attB* are distinguishable from one another not only by the presence of conserved bases in the correct positions but by their absence in the incorrect positions. The 4 half sites are shown, with conserved bases denoted as in part (A). Above each pair is the DBD2-specific consensus, shifted as it would appear in the opposite type of half-site. Bases that do not match the shifted consensus are highlighted in pink. For example, *attB* sites do not have a GT sequence in the position where the conserved GT motif is found in *attP* sites, and *attP* sites do not have a GT motif in the position where it is found in *attB* sites.

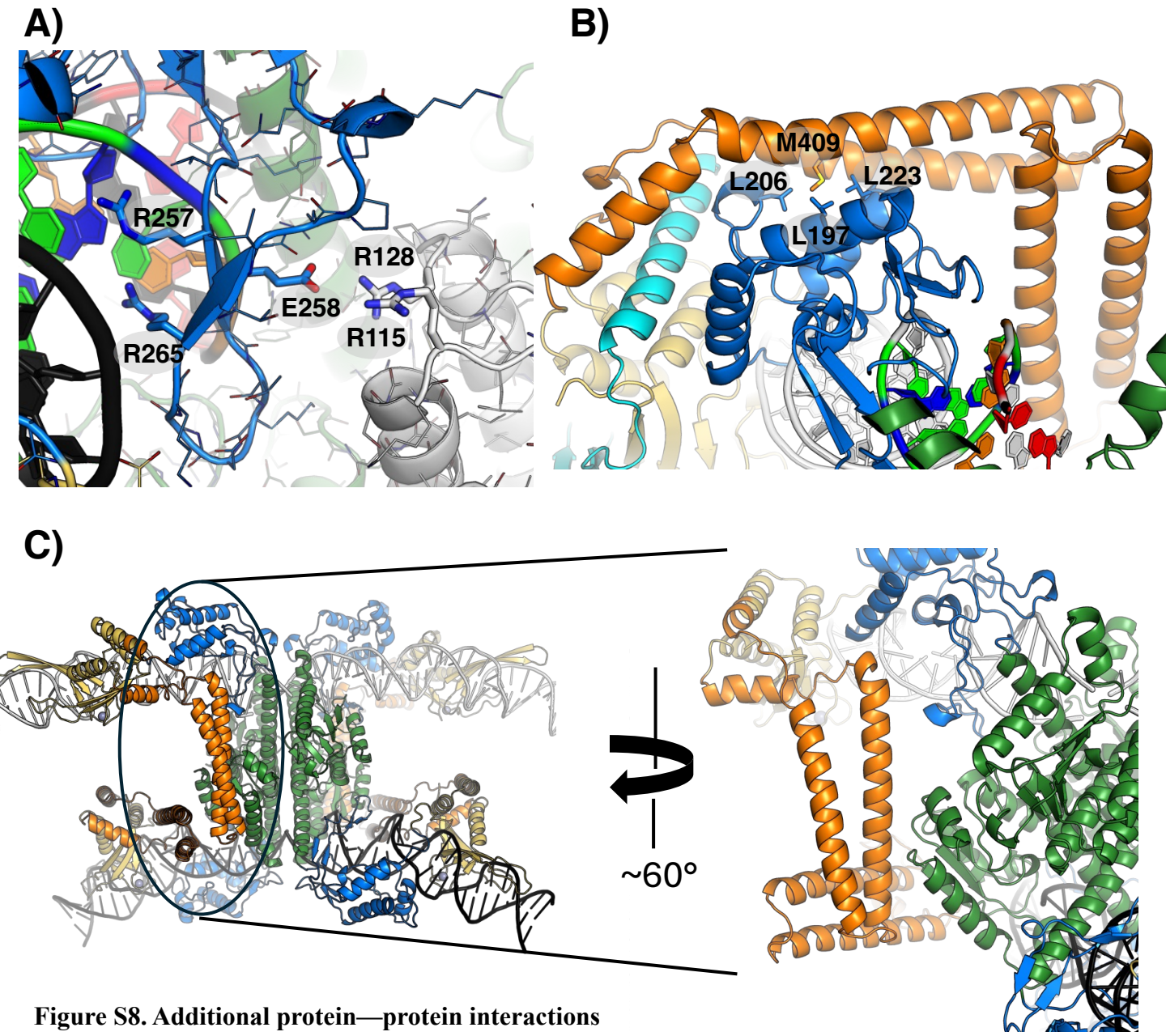


Figure S8. Additional protein—protein interactions

A. The DNA-interacting loop of DBD1 (blue) interacts with the cat domain (in gray). Similar interactions were noted in both in the presence and absence of the RDF and on both types of half-sites.

B. In the presence of the RDF, the *attP* – bound DBD1 interacts with the CC motif (orange) and RDF (cyan). These interactions were not seen on *attB* sites or without the RDF.

C. In the absence of the RDF, the CC motifs interact with the cat domains (green).

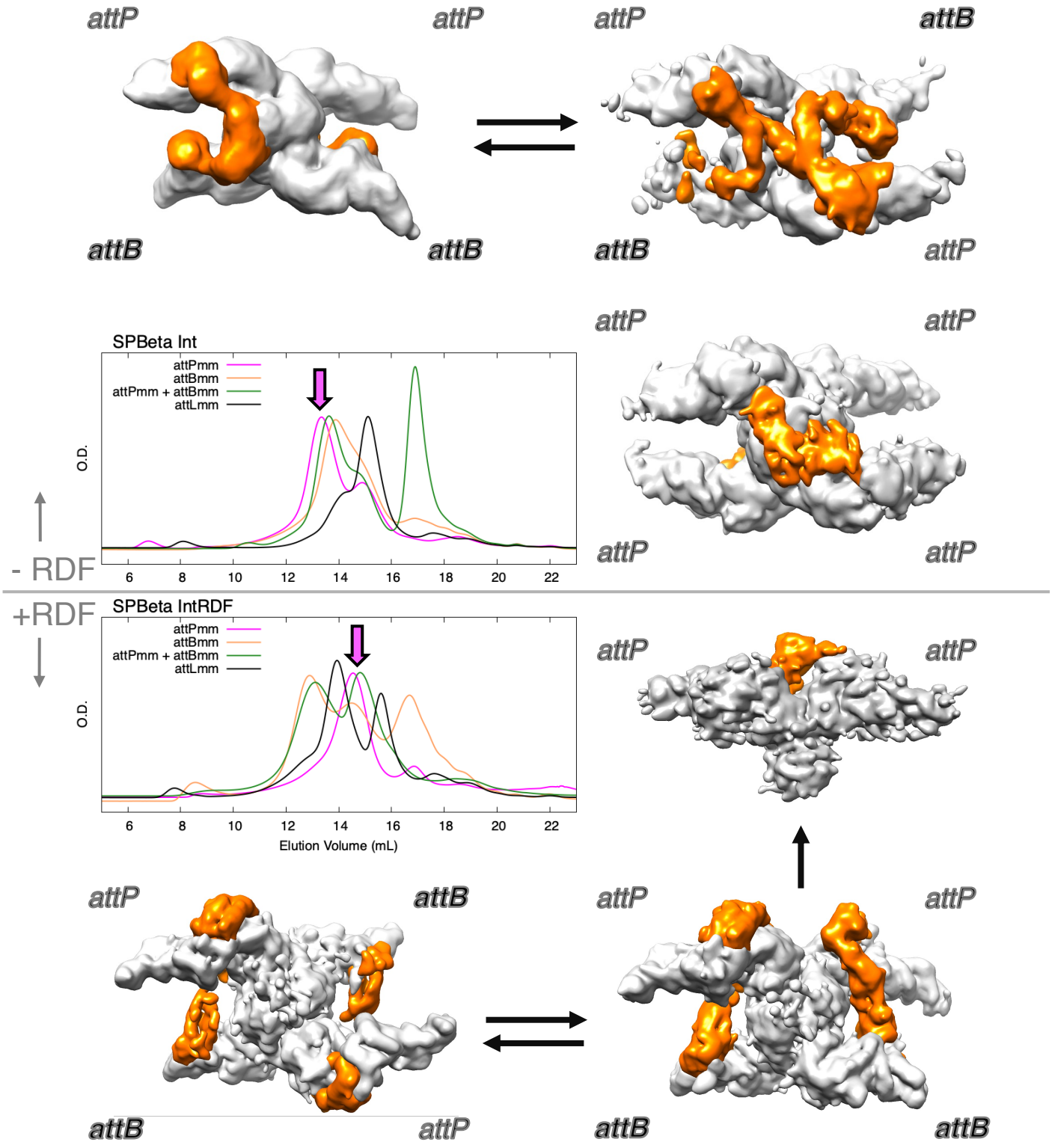
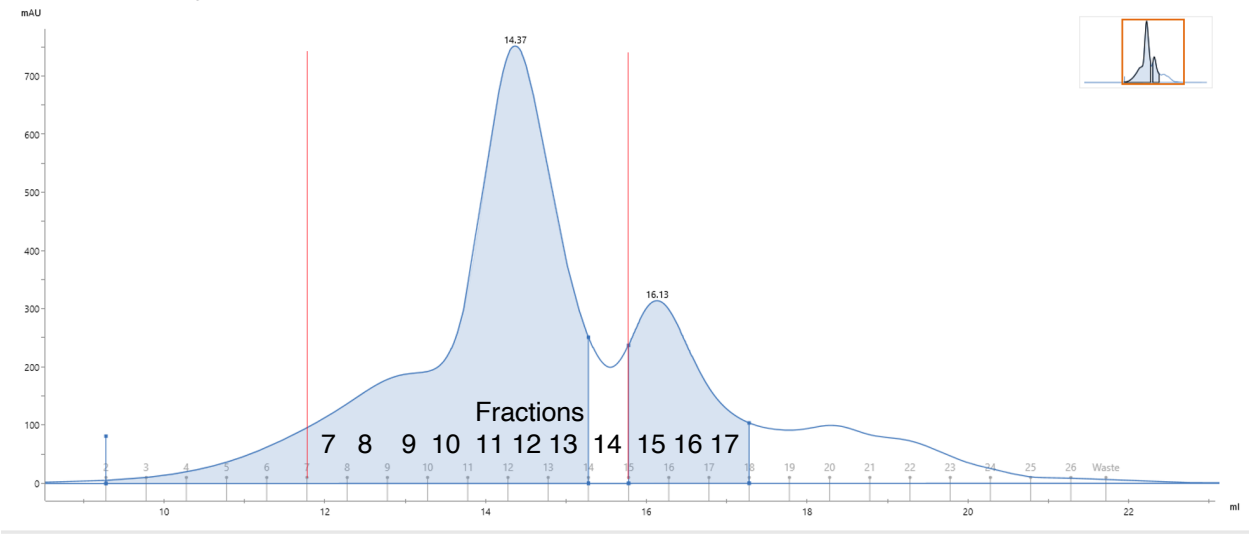


Figure S9. CC-motif handshake interactions regulate directionality during DNA recombination
CC-motifs (in orange) are highlighted in various synaptic complexes.
Elution profiles of the trapped intermediate complexes analyzed via SEC. Top panel: SPbeta integrase complexes in the absence recombination directionality factor (RDF). Bottom panel: SPbeta integrase complexes in the presence of RDF. Note that tetramers synapsing two copies of *attP* were only seen in the absence of the RDF (pink arrows).

A)



B)

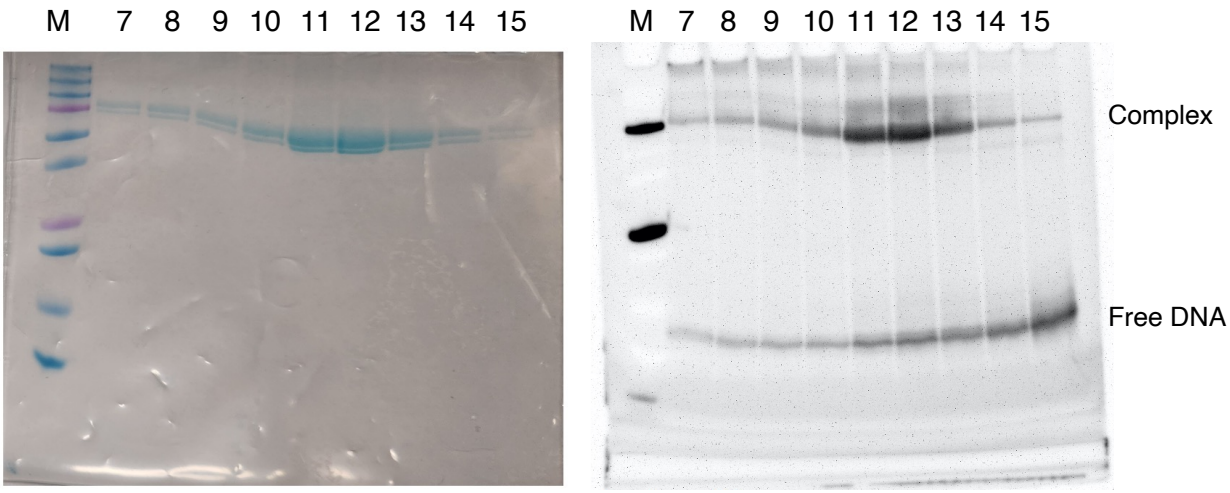


Figure S10. Purification of SPbeta Int-RDF fusion – *attL*mm complex used for cryoEM.

A. Elution profile of *attL*mm bound SPbeta Int-RDF fusion from size exclusion chromatography (SEC).
B. SDS-PAGE of fractions after GelCode blue staining (on the left) and ethidium bromide staining (on the right). Fractions 11 and 12 were pooled to be on EM grids.

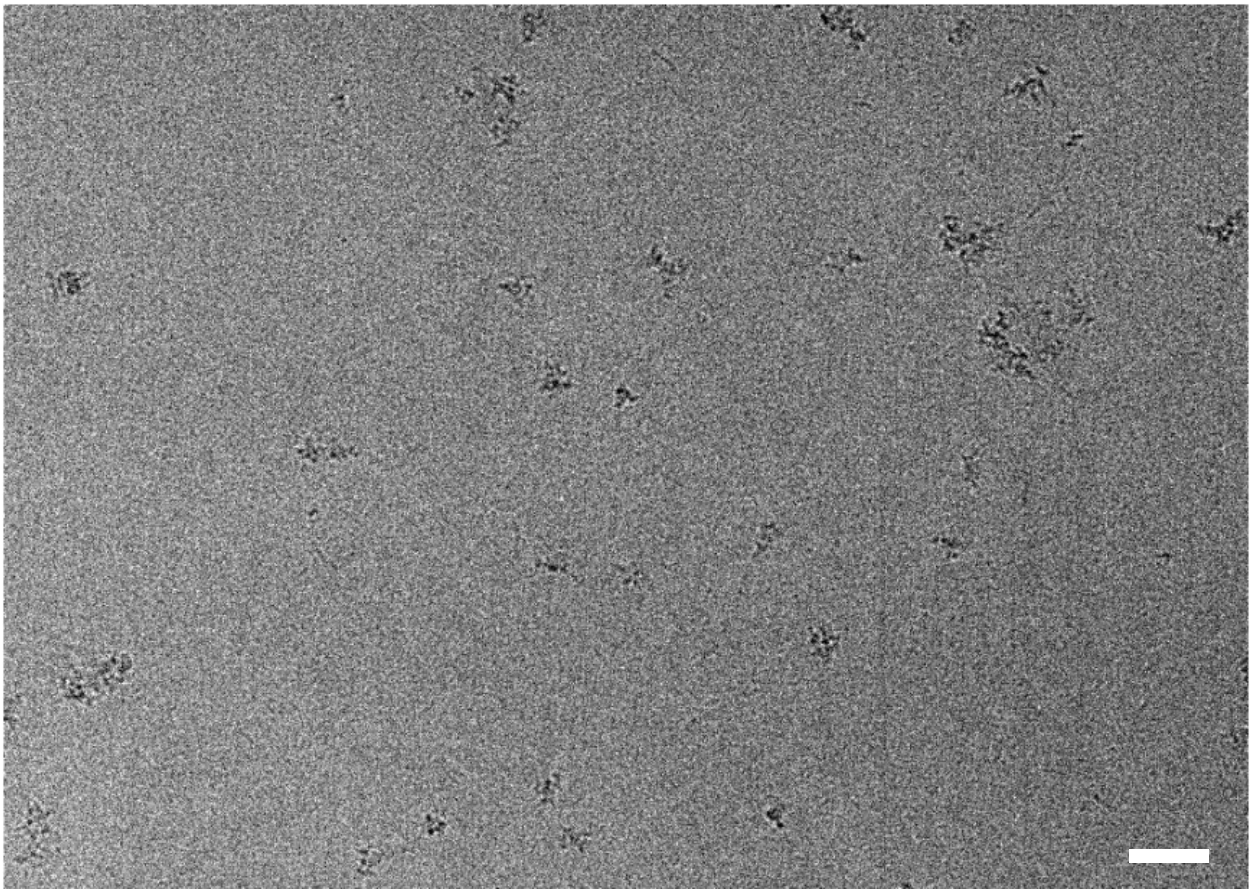


Figure S11. Raw cryo-EM images of SPbeta int-RDF fusion *attL*mm complex.
A raw image of complex particles at 81,000 magnification (1.065Å pixel size) with Titan Krios G3i (Thermo Fisher Scientific). Scale bar corresponds to 340Å.