# The conserved ComP structure with two distinct DNA binding modes - systematic modeling with AlphaFold3, RoseTTAFold2NA, Chai-1 and re-docking in HADDOCK

Stian Aleksander Helsem[1], Kristian Alfsnes[2], Stephan A. Frye[3], Alexander Hesselberg Løvestad[1,4] and Ole Herman Ambur[1]*

[1]Department of Life Sciences and Health, Faculty of Health Sciences, OsloMet, Norway

[2] Division for Infection Control, Department of Bacteriology, Norwegian Institute of Public Health, 0213 Oslo, Norway

[3] Division of Laboratory Medicine, Dept. Microbiology, Oslo University Hospital, Norway

[4]Department of Virology, Norwegian Institute of Public Health, Oslo, Norway

*Correspondence: olam@oslomet.no

Keywords: Competence protein, protein-DNA binding mode, minor pilin ComP, DNA Uptake Sequence, natural competence, transformation, interface, docking, AlphaFold3, RoseTTAFold2NA, Chai-1, HADDOCK

# Abstract

The competence protein ComP is the extracellular DNA binding protein involved in the uptake of DNA in *Neisseria gonorrhoeae* and related Neisserieaceae bacteria. Details of the DNA binding mode of ComP remains enigmatic. Here we identify the ComP orthologs in a set of Neisseriaceae family members, model their common structural domains and their interaction with different preferred binding motifs, DNA Uptake Sequences (DUS). Through systematic in-silico modeling and comparisons using AlphaFold 3, RoseTTAFold2NA, and Chai-1, we bring a new understanding of the interactions between DNA and ComP, and particularly between DUS and ComP. We report six distinct binding modes of which two, here named Epsilon and Gamma, were repeatedly top-scoring models across platforms and different ComPs. The characteristics and robustness of the predicted models from each tool are assessed and discussed. This work may further expand the knowledge on the ComP-DUS interaction and guide further systematic and experimental characterization of these complexes which form extracellularly.

# Introduction

The development of deep learning protein structure prediction algorithms is now allowing the study of protein function and interactions in a detail never seen before. AlphaFold3 (Abramson et al., 2024), RoseTTAFold2NA (Baek et al., 2024), RosETTAFold-All-Atom (Krishna et al., n.d.), NeuralPLexer (Qiao et al., 2024), DiffDock (Corso et al., 2022) and more have led the way to model protein interactions with other proteins, ligands and nucleic acids. Recently Chai-1 was released based with single model training and new functionality (Chai Discovery, 2024), and new implementations of AlphaFold 3 are continuously being developed, with the release of HelixFold3 (Team, 2024) and Boltz-1 (Wohlwend et al., 2024)

Natural transformation is the uptake and recombination of DNA and is found in many Gram positive and Gram negative bacteria (reviewed in Dubnau & Blokesch, 2019). Transformation in the Neisseriaceae family containing the well-studied human pathogens *Neisseria gonorrhoeae* and *N. meningitidis*, is enhanced when short 10-12 nucleotide DNA binding motifs are present in the transforming DNA (reviewed in Ambur et al., 2016). These motifs are named DNA Uptake Sequences (DUS) and are genomically enriched in permissive parts of the genome (Treangen et al., 2008). Eight dialects of DUS have been described, all of which have a conserved 5’-CTG-3’ core essential for transformation (Frye et al., 2013). DUS-dependent transformation provides means to acquire homologous DNA for allelic shuffling (Ambur et al., 2016).

ComP is the minor type IV pilin identified as the DUS receptor (Cehovin et al., 2013) which enhances DUS-dependent transformation and also controls transformation in the absence of

DUS (Berry et al., 2013). ComP is integrated into the extracellularly protruding pilus structure together with the major pilin PilE and other minor pilins (Berry et al., 2016). ComP has a complex globular domain facing the extracellular environment where DNA can bind. The high-resolution 3D structures of ComP (*N. meningitidis* and *N. subflava*) have been reported and were based on crystallography of soluble ComP-maltose-binding-protein fusion proteins and NMR of 6-His tagged ComP (Berry et al., 2016). Their highly similar structures were characterized as containing an N-terminal α-helix, five antiparallel β-strands, the α1-β1- and β1-β2-loops, a DD-region, and two disulfide bridges (Cys-Cys). These disulfide bridges were found crucial for DUS-binding in EMSA-shift assays (Cehovin et al., 2013). ComP has an electropositive stripe, and individual substitutions of three positively charged amino acids therein (R78A, K94A, and K108A) were shown to affect binding and/or transformation negatively (Cehovin et al., 2013). Both *N. meningitidis* and *N. subflava* ComPs have been demonstrated to have higher affinity for DUS containing DNA compared to other sequences using NMR and transformation experiments (Berry et al., 2013, 2016; Cehovin et al., 2013). NMR spectra showed that the DD region, the tip of the ß1-ß2 loop and part of the α1-ß1 loop were involved in DUS binding and were modeled to establish contacts with bases in successive grooves of the dsDNA (Berry et al., 2016). Overexpression of N. $sub_{ComP}$ in a *N. meningitidis* model has been shown to increase non-specific transformation almost up to the level of DUS-mediated transformation (Berry et al., 2013) showing that ComP has a DUS-independent positive effect on transformation. ComP has been found to preferentially mediate transformation of dsDNA, although single-stranded reverse strand DUS shows some transforming ability in a *N. gonorrhoeae* model (Duffin & Seifert, 2012).

All major structural work on the ComP-DUS interaction has focused on the ComP proteins of *N. meningitidis* and *N. subflava* (Berry et al., 2013, 2016; Cehovin et al., 2013).

Fewer efforts have been put into studying the modeled ComP-DUS binding mode. Berry et al. (2016) and Hughes-Games (2020) performed docking analyses on *N. subflava* NJ9703$_{ComP}$ and *N. meningitidis* 8013$_{ComP}$. Their molecular docking results both show that the α1-β1 and β1-β2 loops interact with major groove and the DD-region with minor groove of the DNA.

To our knowledge no systematic modeling of ComP with DUS containing DNA has been performed using the newly available deep-learning based platforms. AlphaFold 3 (AF3), RoseTTAFold2NA (RF2NA) and Chai-1 represent the state-of-the-art of modeling protein-nucleic acid (protein-NA) complexes and hence may provide means to better understand the ComP-DUS interaction. Using five of the DUS dialects from Frye et al. (2013) as the basis for comparatively modeling ComP-DNA/DUS interactions using modeling algorithms, this study brings a new understanding of the ComP-DUS interaction.

# Materials and methods

## Data material

ComP orthologs were obtained using a series of BLAST searches (Camacho et al., 2009), using the structures with PDB accession 2NBA (Berry et al., 2013), 2M3K (Cehovin et al., 2013) and 5HZ7 (Berry et al., 2013) as queries against the RefSeq non-redundant database, specifying the Neisseriaceae clade as target. ComP amino acid sequences from seven Neisseriaceae species modeled with DNA containing their native 12 bp DUS dialects were used in this study. The ComP-DUS pairs were: *Eikenella corrodens* ATCC 23834 ComP-AG-eikDUS (E. cor$_{nat}$), *Neisseria meningitidis* MC58 ComP-AT-DUS (N. men$_{nat}$), *Kingella denitrificans* ATCC 33394 ComP-AA-king3DUS (K. den$_{nat}$), *Neisseria mucosa* ATCC 25996 ComP-AG-mucDUS (N. muc$_{nat}$), *Bergeriella denitrificans* ATCC 33394 ComP-AG-DUS (B. den$_{nat}$), *Neisseria cinerea*

ATCC 14685 ComP-AT-DUS (N. $cin_{nat}$) and *Neisseria subflava* NJ9703 ComP-AG-DUS (N. $sub_{nat}$). These will colloquially be called native ComP-DUS complexes. When referring to the strain-specific ComPs and their native DUS, the suffix nat will be appended to the species name (for example N. $sub_{nat}$). The five DUS dialects and their native ComPs were previously reported by Frye et al. (2013). Refer to Table S1 for GenBank/RefSeq ComP accession numbers. The truncated ComP amino acid sequences (28 amino acids trimmed off the N-terminal part of the mature $ComP_{N.\ men}$ as in published structures) and nucleotide sequence of DUS-dialects used, can be accessed in Table S1.

**Modeling of ComP-DUS complexes**

To model native ComP-DUS complexes, ComPs were modeled with their native 12 bp DUS (dsDNA) in AF3 (web server), RF2NA (local installation) Chai-1 v. 0.2.0 (local installation). A minimum of 20 native AF3 replicates (33 for B. $den_{nat}$ and 37 for N. $sub_{nat}$) and 100 native runs in RF2NA and Chai-1 were performed for each $ComP_{nat}$. Customization of multiple sequence alignment (MSA) input to Chai-1 allowed for different MSA schemes (Table S2), using HHblits v. 3.0.0 (Steinegger, Meier, et al., 2019) to search the Big Fantastic Database (BFD) (Jumper et al., 2021; Steinegger, Mirdita, et al., 2019; Steinegger & Söding, 2018) and Uniref30 (Mirdita et al., 2022) databases for ComP MSAs, using the same HHblits parameters as in RF2NA. DUS MSAs were created based on genomic DUS counts in selected genomes (not shown). 10 replicates were run for each MSA scheme. This generated a grand total of 3500 mmCIF/PDB models in Chai-1 (500 per $ComP_{nat}$), ~850 models in AF3, with more than 100 models per $ComP_{nat,}$ and 700 models in RF2NA, with 100 models per $ComP_{nat}$. Criteria for assigning successful models were ipTM > 0.6 in AF3 and Chai-1 and PAE $\leq$ 10 in RF2NA.

To investigate AF3's ability to correctly identify native ComP-DUS complexes, as measured by CPPM, we carried out modeling using scrambled DUS sequences in addition to the native ones. ComPs were modeled with scrambled 12 bp DUS dialects in AF3, not allowing for the conserved trinucleotide 5'-CTG-3' nor tetrahomomers (ie. "AAAA") in the sequence, while containing the same nucleotide content as the native DUS. Wilcoxon rank sum significance tests were then performed to determine whether runs with native DUS yielded higher confidence metrics (see below) than scrambled DUS runs.

To evaluate the modeling reproducibility of the different binding modes while also assessing the docking approaches used in previous studies (Berry et al., 2016; Hughes-Games, 2020) (Berry et al., 2016; Hughes-Games, 2020), we performed re-docking experiments in HADDOCK (João M. C. Teixeira et al., 2024). The top ranking models of $ComP_{nat}$ and DUS from RF2NA were re-docked. Following the suggestion by (Esmaeeli et al., 2023), prior to molecular docking in HADDOCK, the first 10 all-atom non-trivial normal modes with lowest frequency for both the ComPs and DUS were generated using the R module bio3d (Grant et al., 2006, 2021; Skjærven et al., 2014, 2016). Selected normal mode conformations were fed into HADDOCK as ensemble PDB files and cross-docked (all against all), generating 400.000 rigid bodies per $ComP_{nat}$, of which, after flexible and molecular dynamics refinement, the top 200 models (HADDOCK score) were assigned binding mode. For N. $sub_{nat}$, all solvent-accessible ComP residues that showed a NMR CSP shift below DNA concentration of 20 mM in (Berry et al., 2016) were specified as ambiguous restraints. For the other $ComP_{nat}$, the same residues were used as ambiguous restraints only if conserved in the ComP alignment (Fig. S2)

The following measures were used to assess the quality of predicted models from Chai-1, AF3, RF2NA and HADDOCK:

- Predicted interface template modeling score (ipTM; AF3 and Chai-1; confidence range 0-1, where >0.6 (grey zone to high quality) were considered successful models.
- Predicted aligned error (PAE; AF3, Chai-1 and RF2NA; unit Å; lower values indicate higher confidence in structure and values and <10 Å were considered successful RF2NA models). Significant differences between models are reported.
- Predicted local distance difference test (pLDDT; AF3, Chai-1 and RF2NA; confidence range 0-100). Significant differences between models are reported.
- Chain pair PAE minimum (CPPM; AF3; Å; measure used to distinguish probable binders from non-binders in interactions between chains). Significant differences between models are reported.
- HADDOCK score (HADDOCK; weighted sum of various energy terms, including van der Waals forces, electrostatics and desolvation energy)
- DockQ - continuous quality measure of similarity (confidence range 0-1) for docking models contained in two different PDB/mmCIF files used herein to assess consistency in predicted ComP-DUS models within and between AF3, Chai-1 and RF2NA (Basu & Wallner, 2016; Mirabello & Wallner, 2024).

All Chai-1 and AF3 models having ipTM $\geq$ 0.6 and RF2NA models having PAE $\leq$ 10 (RF2NA does not output ipTM) were deemed robust and visually inspected in PyMOL v. 3.0.0 and binding mode assigned according to the descriptions below. In ComP-DNA models for which ipTM never reached a 0.6 threshold, all models were manually inspected

and assigned binding mode. Based on similarity across models, we define the following binding modes:

γ (Gamma): α1-β1 and β1-β2 loops interact with minor groove and DD-region with major groove

δ (Delta): α1-β1, β1-β2 and DD-region all interact with the minor groove

ε (Epsilon): α1-β1 and β1-β2 loops interact with major groove and DD-region with minor groove

ζ (Zeta): α1-β1, β1-β2 and DD-region all enter the major groove

η (Eta)*: β4- and β5 sheets interact with the major groove and the N-terminal end of ComP is in close proximity of either the start or end of DUS (exclusively seen in HADDOCK K. $den_{nat}$, B. $den_{nat}$ and N. $men_{nat}$)

θ (Theta): α 1-β , β1- β2 and DD-region all enter the major groove of a partially unwound DNA helix (exclusively seen in Chai-1 models of N. $cin_{nat}$)

ο (Other): Did not clearly fall into any of the previous.

* Part of ο (Other) in Table 1

Representative examples of ComP-DUS complexes in the different binding modes are shown in Figure S1.

**Internal consistency, cross-platform convergence and divergence**

To investigate intra-platform consistency of each AF3, Chai-1 and RF2NA results, all against all DockQ scores were computed for the top-ranking models across replicate runs within the same software. This was devised as a test of platform consistency in predicting the same

binding mode. Additionally, the degree of agreement in the results across AF3, Chai-1 and RF2NA was quantified by calculating all against all DockQ scores for top ranking models. To avoid giving weights to lower ranking models, only AF3 and Chai-1 models having ipTM $\geq$ 0.6 were considered in the cross-platform DockQ comparisons (Abramson et al., 2024; Zhang & Skolnick, 2004), except for N. $men_{nat}$ and N. $cin_{nat}$ in AF3 and N. $muc_{nat}$ in Chai-1, for which ipTM $\geq$ 0.6 was never reached (all models were included for these cases). Likewise, only those RF2NA models having mean interface predicted aligned error (PAE) < 10 were considered (Baek et al., 2024).

# Results

### AF3 ComP structure modeling

Here, using AF3, we explored the structures of the same two ComP ptoteins (N. $men_{ComP}$ and N. $sub_{ComP}$) as previously reported (Berry et al., 2016; Cehovin et al., 2013) and five additional orthologous proteins from other Neisseriaceae members (*B. denitrificans*, *E. corrodens*, *N. mucosa*, *N. cinerea* and *K. denitrificans*) (Fig. 1).

**Figure 1.** Overlay of the best AF3 models of ComP for seven species, aligned on the α-carbons of the model with the best ipTM score, K. $den_{ComP}$ Purple. Green: N. $sub_{ComP}$. Turquoise: B. $den_{ComP}$. Yellow: N. $muc_{ComP}$. Pink: E. $cor_{ComP}$. Grey: N. $men_{ComP}$. Blue: N. $cin_{ComP}$. The β-strands are numbered and domains labelled.

As shown in Figure 1, the alignment of the AF3 models with highest ipTM score on the α-carbons of K. $den_{ComP}$ (Fig. 1) revealed remarkable similarity across all seven ComPs and that all contain the same domains as previously described (Berry et al., 2016; Cehovin et al., 2013). Identifiable and ubiquitous domains are an N-terminal α-helix, four antiparallel β-strands, α1-β1- and β1-β2-loops, the DD-region, and two essential disulfide bridges (not shown) (Berry et al., 2016). Additionally, a fifth β-strand was consistently predicted in AF3. Minor differences were seen across the ComPs, including prediction of a short, additional α-turn embedded in the β1-β2-loop in B. $den_{ComP}$ and E. $cor_{ComP}$ and an α-turn located in the DD-region in N. $muc_{ComP}$.

During modeling ComP-DNA we observe minor structural variation in different domains (mainly DD-region and β1-β2 loop) for the same ComP across different replicate runs, even within the same platform. In the ComP-DUS modeled structures in Figure 2, these varying regions can be seen as imperfectly aligned domains. We do not present ComP structure overlays from Chai-1 and RF2NA beyond those in Figure 2. We note however that RF2NA generally produced models that lack the disulfide bridge connecting the β1-sheet and the DD-region despite these being present in all three ComP PDB structures (2NBA, 5HZ7 and 2M3K) used as templates by RF2NA. We interpret that the absence of this disulfide bridge could allow more flexibility of the DD-region and potentially more substantial structural alterations. In a smaller fraction of the RF2NA models, also the disulfide bridge connecting β4- and β5-sheets were lacking which could affect the integrity of the β-sheet.

**Modeling native ComP-DUS complexes**

All top native ComP-DUS models in RF2NA, Chai-1 and AF3 are shown in Fig. 2, each aligned on the α-carbons of the ComP of the top model. The following binding modes were predicted by the three platforms for the seven different native ComP-DUS complexes.

**Figure 2**. The best ranking models from RF2NA, Chai-1 and AF3 for all seven native ComP-DUS complexes, aligned on the α-carbon of the models with the top ipTM score. The platforms are labelled in yellow and $ComP_{nat}$ and the confidence measures in white. **A**: $N.sub_{nat}$; **B**: $B.den_{nat}$; **C**: K. $den_{nat}$; **D**: N. $muc_{nat}$; **E**: $E.cor_{nat}$; **F**: N. $cin_{nat}$; and **G**: N. $men_{nat}$. * ipTM never reached 0.6

**N. $sub_{nat}$:** All successful models by AF3 (158/158) and Chai-1 (428/428) were in Epsilon, while 45/46 successful RF2NA models were observed in Delta and 1/46 in Gamma. Overall, the Chai-1 models were found to be significantly better than AF3 models, while both Chai-1's and AF3's models were significantly better than RF2NA's (both PAE and pLDDT) (Code Output S1). Note the consistently flipped out base (T; pos +5 in Frye et al., 2013 numbering) in the Chai-1 models. Some of the Chai-1 models for $N.sub_{nat}$ had the highest ipTM scores (≈ 0.84) observed for all native ComP-DUS models.

**B. $den_{nat}$:** All successful models by AF3 (165/165) and Chai-1 (91/91) were Epsilon, as were the majority (20/22) of the successful RF2NA models. RF2NA modeled a single Gamma and one Other. The AF3 models were found to be significantly better in terms of PAE than those from Chai-1, which again were significantly better than the RF2NA models. pLDDT was not significantly different between AF3 and Chai-1, but both were when compared to RF2NA.

**K. $den_{nat}$**: All successful models by AF3 (105/105) and Chai-1 (446/446) were Epsilon. Successful models from RF2NA showed high variation in binding mode with significantly lower quality than AF3 and Chai-1 (PAE and pLDDT). Of these, 69/98 were robust models in Zeta, 25/98 Epsilon and 4/98 Gamma. The Chai-1 and AF3 models showed no significant difference in terms of PAE, albeit the AF3 models had significantly higher pLDDT.

**N. muc$_{nat}$** : All successful models by AF3 (105/105) were Epsilon. No Chai-1 models reached ipTM $\geq$ 0.6, and all of the 500 best scoring models were Epsilon. Successful RF2NA models were Gamma (10/11) and one Other. The models from both Chai-1 and AF3 were significantly better than those from RF2NA in terms of PAE, but pLDDT was not significantly different between Chai-1 and RF2NA. PAE was significantly better in AF3 than Chai-1. On the contrary, and despite the below cut-off ipTM scores in Chai-1, pLDDT was significantly better in Chai-1 than AF3.

**E. cor$_{nat}$**: All successful models by AF3 (95/95) were Gamma. Successful Chai-1 models were ambiguous in terms of binding mode in that the models with lowest ipTM were either Gamma (138/381) or Other (86/381), while the better models (157/381) were Epsilon. The successful RF2NA models were also ambiguous in terms of binding mode, with Gamma (26/81) Epsilon (23/81), Zeta (25/81), Delta (1/81) and Other (6/81) assignments. The Chai-1 models were significantly better than both AF3 and RF2NA.

**N. cin$_{nat}$**: The AF3 models never reached ipTM $\geq$ 0.6 and of the 100 best scoring models, 96/100 were in Gamma and 4/100 in Epsilon. In Chai-1, we observed Theta for all successful models (395/395), a mode exclusively seen in Chai-1 for this ComP$_{nat}$. In these Theta models, there was a complete strand separation in one end of the DNA and one flipped out base (T; pos +7 following numbering in Frye et al., 2013). The successful RF2NA models were either Gamma (40/52) or Delta (12/52). Both AF3 and Chai-1 models were significantly better than RF2NA in terms of PAE and pLDDT, while Chai-1 was significantly better than AF3.

**N. men$_{nat}$**: The AF3 models never reached ipTM $\geq$ 0.6, and of the 100 best scoring models, 76/100 were Gamma, 21/100 Epsilon and 3/100 Delta. The successful Chai-1 models were in Epsilon (12/13) and Zeta (1/13). The RF2NA models were in Gamma (56/91) and Delta (35/91). Chai-1 gave significantly better PAE and pLDDT scores than both AF3 and RF2NA.

Interestingly, the RF2NA models were of significantly better pLDDT than the AF3 models, albeit having significantly weaker PAE.

For the re-docking of top RF2NA models for each $ComP_{nat}$, we observed mostly Epsilon and Gamma, further solidifying the support for these binding modes. However, because the input DNAs were highly distorted (due to normal mode calculation), it was often challenging to discern minor from major groove and assign binding mode. This was also why the HADDOCK output models were not used in the cross-platform DockQ analyses. Despite this, for the re-docks where the DNA was close to its input B-DNA state, many models were classical Epsilon and Gamma. Thus, even though the categorization of binding modes for HADDOCK models must be cautiously considered, we assign the binding following HADDOCK mode distributions and mode bias (strong bias when above 70% of assigned modes) for each $ComP_{nat}$:

**N.sub$_{nat}$:** 159/200 Gamma, 25/200 Epsilon and 15/200 “Other”, strong bias Gamma.

**B.den$_{nat}$:** 145/200 Epsilon, 54/200 Eta (“Other”) and 1/200 Gamma, strong bias Epsilon.

**K. den$_{nat}$**: 200/200 Epsilon, unambiguously Epsilon.

**N. muc$_{nat}$**: 145/200 Epsilon, 53/200 Gamma and 2/200 “Other”, strong bias Epsilon.

**E.cor$_{nat}$**: 170/188 models are Epsilon and 18/188 Zeta, strong bias Epsilon.

**N. cin$_{nat}$**: 121/200 Gamma and 79/200 Epsilon, bias Gamma.

**N. men$_{nat}$**: 130/200 models are Epsilon, 68/200 Gamma and 2/200 Zeta, bias Epsilon.

All mode assignments are summarized in Table 1.

**Table 1.** Across platform and $ComP_{nat}$ binding mode predictions with ipTM $\geq$ 0.6 (AF3 and Chai-1) or PAE $\leq$ 10 (RF2NA). **ε** = Epsilon, ***ζ*** = Zeta, ***γ***=Gamma, **δ**=Delta and **o**=other (incl. Eta and Theta)

| Platform | AF3 | | | | | Chai-1 | | | | | RF2NA* | | | | | HADDOCK | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mode/ $ComP_{nat}$ | **ε** | ***ζ*** | ***γ*** | **δ** | **o** | **ε** | ***ζ*** | ***γ*** | **δ** | **o** | **ε** | ***ζ*** | ***γ*** | **δ** | **o** | **ε** | ***ζ*** | ***γ*** | **δ** | **o** |
| $N.sub_{nat}$ | 258 | - | - | - | - | 428 | - | - | - | - | - | - | 1 | 45 | - | 26 | - | 159 | - | 15 |
| $B.den_{nat}$ | 165 | - | - | - | - | 91 | - | - | - | - | 20 | - | 1 | - | 1 | 145 | - | 1 | - | 54 |
| $K.\ den_{nat}$ | 146 | - | - | - | - | 445 | - | - | - | - | 25 | 69 | 4 | - | - | 200 | - | - | - | - |
| $N.muc_{nat}$ | 105 | - | - | - | - | *500 | - | - | - | - | - | - | 10 | - | 1 | 145 | - | 53 | - | 2 |
| $E.cor_{nat}$ | - | - | 95 | - | - | 157 | - | 138 | - | 86 | 23 | 25 | 26 | 1 | 6 | ‡170 | ‡18 | - | - | - |
| $N.cin_{nat}$ | 4 | - | *96 | - | - | - | - | - | - | †395 | - | - | 40 | 12 | - | 79 | - | 121 | - | - |
| $N.men_{nat}$ | *21 | - | *76 | *3 | - | 12 | 1 | - | - | - | - | - | 56 | 35 | - | 130 | 2 | 68 | - | - |
| SUM | 699 | 0 | 267 | 3 | 0 | 1633 | 1 | 138 | - | 481 | 68 | 94 | 136 | 93 | 7 | 895 | 20 | 402 | - | 69 |

*ipTM < 0.6 or ipTM not available (RF2NA)

† All in Theta mode

‡ For E. $cor_{nat}$ in HADDOCK, 200 docking solutions was not reached and 188 final docking solutions were found

The only two modes predicted by all three platforms and HADDOCK were Epsilon (3295/5005) and Gamma (953/5005). Generally, AF3, Chai-1 and HADDOCK predicted more Epsilon and RF2NA more Gamma. The four N.sub$_{nat}$, B.den$_{nat}$, K. den$_{nat}$ and N. muc$_{nat}$ were more robustly predicted in Epsilon than E. cor$_{nat}$, N. cin$_{nat}$ and N. men$_{nat}$, which were more inclined to be predicted in Gamma. The Zeta mode (115/5005), which resembles Epsilon, was not modeled by AF3 and mostly by RF2NA. The Delta mode (93/5005), which resembles Gamma, was not modeled by Chai-1 and mostly by RF2NA. Other modes were not modeled by AF3. The unique Theta mode (359/5005) was only modeled by Chai-1. HADDOCK found strong general support for the Epsilon and Gamma modes and deviate from the modeling platforms most notably in assigning Gamma for N.sub$_{nat}$ (only 1 recorded in RF2NA), a strong Epsilon bias for E. cor$_{nat}$ and a bias for Epsilon for N. men$_{nat}$.

**Internal consistency and cross-platform convergence**

When assessing the internal consistency of AF3, Chai-1 and RF2NA native ComP-DUS models using DockQ scores, both AF3 and Chai-1 showed higher model consistency across replicate runs than RF2NA. AF3 showed a significantly higher median DockQ score (0.841) compared to the other two platforms (Chai-1: 0.680; RF2NA: 0.617), as determined by pairwise Wilcoxon rank-sum tests (Code Output S3). Also, Chai-1 showed a significantly higher median DockQ score compared to RF2NA.

In order to explore cross-platform convergence for each ComP$_{nat}$, we assessed structural similarity of predicted models from all platforms directly using DockQ scores. The DockQ scores were categorized as higher or lower according to the 0.23 (low), 0.49 (intermediate) and 0.8 (high) limits (Figure 3).

**Figure 3.** Violin plots on DockQ scores for cross-platform model comparisons. The median and interquartile ranges are shown as horizontal bars, and model comparisons are depicted as swarms inside the violins. Horizontal red, yellow and green dashed lines represent DockQ of 0.23 (low), 0.49 (intermediate) and 0.8 (high), respectively. Red violins represent Epsilon, blue Gamma, light grey Other, dark grey Delta and green pairs Different (models assigned to different binding modes).

**N. sub$_{nat}$**: The cross-platform model comparisons showed that the Epsilon binding mode seen in some Chai-1 and AF3 models were virtually identical, reaching DockQ $\geq$ 0.8 (Fig. 3). The bulk of AF3 vs. Chai-1 model pairings were in the medium range (DockQ≈0.6), and all were in Epsilon, showing high cross-platform convergence. Both categories comparing to RF2NA had the majority of pairings in the acceptable range, and none were in the same binding mode ("Different").

**B. den$_{nat}$**: All AF3 vs. Chai-1 model pairings were in the medium range and in Epsilon, with some pairings being near high-quality. All AF3 vs. RF2NA pairings were in the acceptable range, except for a residual of pairings that were in the lower medium range, being either in Different or Epsilon binding mode. All Chai-1 vs. RF2NA pairings were in the acceptable range (Different or Epsilon).

**K. den$_{nat}$**: Some model pairings reached high DockQ $\geq$ 0.7 for AF3 vs. Chai-1 comparisons, but most pairings were in the acceptable range. All were in Epsilon. For the comparisons to RF2NA, the majority of model pairings were in the acceptable range, yet some were in the medium range. The distributions of different and Epsilon pairings were very similar.

**N. $muc_{nat}$**: Some models reached high DockQ scores close to 0.8 for AF3 vs. Chai-1, with a also broad range of DockQ scores, both acceptable and medium. The comparisons to RF2NA were mostly in the acceptable range, represented by both Different and Epsilon modes.

**E. $cor_{nat}$**: Gamma was the most often co-occurring binding mode in AF3 vs. Chai-1, with pairings of both medium and acceptable DockQ scores. Some acceptable model pairings were assigned to different modes. In the RF2NA comparisons, the majority of pairings were in the acceptable range, both Gamma, Epsilon and Different being represented, in addition to a binding mode that did not fit into any of the described ones ("Other").

**N. $cin_{nat}$**: AF3 vs. Chai-1 comparisons never reached acceptable DockQ, and all pairings were in different binding modes. However, AF3 vs. RF2NA pairings reached high-quality DockQ for the Gamma mode, albeit with most pairings in the medium range. Some models were in different binding modes. All Chai-1 vs. RF2NA comparisons were below the acceptable range.

**N. $men_{nat}$**: Most AF3 vs. Chai-1 pairings were in the acceptable range, being Different or Epsilon, with some pairings in the medium range. AF3 vs. RF2NA have both Delta, different and Gamma represented in the acceptable and medium ranges, with the two latter having some pairings close to the high-quality range. All Chai-1 vs. RF2NA pairings were of different modes and in the acceptable range.

We observe generally higher quality of the AF3 and Chai-1 models relative to RF2NA in terms of both PAE and pLDDT, a consistent trend across binding mode predictions (Fig. 4).

**Figure 4.** Scatter plot of PAE and pLDDT values of $ComP_{nat}$ models from AF3 (red), Chai-1 (blue) and RF2NA (green) for all binding modes observed. Intra-platform sub-clustering reflects different $ComP_{nat}$ models as described in Table 1.

For Epsilon (Figure 4, first panel), the AF3 and Chai-1 models formed clusters of models with PAE range 2.5-6 and pLDDT range 78-93. The strongest scoring models were Epsilon in AF3 followed by Chai-1. The AF3 Epsilon models covered a wider range (strongest and weakest) than Chai-1 Epsilon in regard to both measures. Both AF3 and Chai-1 distributions reflected modes from five not fully overlapping $ComP_{nat}$ each (Table 1). All RF2NA models across modes (no Theta prediction in RF2NA) clustered and were skewed towards considerably higher PAE range 8.5-10 than AF3 and Chai-1 and covered a pLDDT range of 82-85.

Zeta was observed once in Chai-1 (N. $men_{nat}$) with PAE 3.5 and pLDDT 86 and repeatedly in RF2NA (K. $den_{nat}$ and E. $cor_{nat}$) with PAE range 8.3-10 and pLDDT range 81-84. The strongest models that did not clearly fall into any of the categorized binding modes ("Other") still had high pLDDT of 88 and low PAE of 3, while those from RF2NA had PAE 9.5 and pLDDT 84.

Gamma found its strongest models from Chai-1 (PAE 2.5 and pLDDT 87), yet these models were of lower ipTM score than their counterparts in AF3, as outlined in Table 1. The Chai-1 Gamma cluster were all E. $cor_{nat}$. As with Epsilon described above, the AF3 Gamma models covered a wider range than Chai-1 regarding both measures and reflected E. $cor_{nat}$ and an additional two $ComP_{nat}$ (N. $cin_{nat}$ and N. $men_{nat}$). RF2NA was again skewed toward higher PAE and lower pLDDT scores and represent all seven $ComP_{nat}$.

The Delta mode was found only three times in AF3 (N. $men_{nat}$), having PAE range 5-5.5 and relatively low pLDDT range 78-82. This mode was seen repeatedly in RF2NA with PAE range 8.3-10 and pLDDT range 79-85 reflecting models for N. $cin_{nat}$, N. $men_{nat}$ and N. $sub_{nat}$.

The best Theta models, a binding mode only found in N. $cin_{nat}$ Chai-1 models, had PAE 2.5 and pLDDT 87. Theta models clustered in a similar manner as a subset of the Epsilon models.

Finally, all but one (B. $den_{nat}$) Other models in Chai-1 and RF2NA were E. $cor_{nat}$ (Figure 2 and Table 1). The Chai-1 models were again considerably stronger than those of RF2NA.

## AF3 modeling quality in matching native vs. scrambled DUS

In order to explore modeling quality and consistency further, we comparatively modeled native ComP-DUS complexes to scrambled DNA controls and monitored the resulting ipTM and CPPM distributions for all models (Figure 5 A and B). Since the native DUS dialects were the same in the two pairs N. $men_{nat}$/N. $cin_{nat}$ (AT-DUS) and B. $den_{nat}$/N.$sub_{nat}$ (AG-DUS) these were combined in the assessment. AF3 was chosen as platform as it covered most $ComP_{nat}$ (6/7) with robust models. Wilcoxon rank sum tests (Table S3), using significance level $\alpha$=0.05, showed that the ipTM distributions (Figure 5 A) for native DUS models ("Native") and scrambled DUS models ("Scrambled") were significantly different in all comparisons except AA-king3DUS. Furthermore, ipTM in Native was significantly better (higher) than Scrambled for AG-DUS, AG-eikDUS and AG-mucDUS. In contrast, for AT-DUS the ipTM distribution was significantly lower in Native than Scrambled. This suggests a greater difficulty for AF3 to generate high-quality models for Scrambled than Native in all (B. den, N. sub, E. cor, N. muc) except AA-king3DUS (K. den) and AT-DUS (N. cin and N. men). CPPM (5 B) reflects ipTM in that the distributions were significantly different in all comparisons except AA-king3DUS. CPPM was significantly better (lower) in Native than Scrambled for AG-DUS, AG-eikDUS and AG-mucDUS. For AT-DUS, CPPM was significantly higher in Native than Scrambled, corroborating the results for ipTM and reflecting AF3's difficulty in predicting native ComP/AT-DUS complexes and high variation in the models for these complexes as also seen in Fig. 2. This also indicates that AF3 was not distinguishing AA-king3DUS from its scrambled versions.

# Discussion

## ComP structures

In this study, three different deep learning protein structure prediction platforms and one traditional molecular docking platform were used to model the interaction between ComP and DUS, and the output from AF3 allowed us to compare the modeled 3D structures of ComPs. We found that there is high similarity in ComPs overall binding modes and domain organization, structure conservation, showing that all tested proteins are the likely DUS receptors, although their primary sequence diverge. The seven ComPs studied herein were chosen as they represent both a diverse set regarding primary sequence, were from three different genera (*Neisseria/Bergeriella, Eikenella* and *Kingella*) and had established associations to five different DUS dialects (Frye, 2016). Differences in amino acid composition of ComPs, as well as minor differences in structural domains in ComPs, e.g. the here predicted α-turns in AF3 B. $den_{ComP}$, E. $cor_{ComP}$, and N. $muc_{ComP}$, might together help understand general and specific characteristics of the ComP-DUS interaction.

It remains unclear why RF2NA generally did not predict a disulfide bridge between β1-sheet and the DD-region. Possibly the RF2NA algorithm predicts a breakage of the disulfide bond upon interacting with the DNA. However, we deem this to be a biologically unlikely break, as there is no known mechanism by which DNA can induce breakage of disulfide bonds in proteins to our knowledge. Rather, this is achieved using either electron capture dissociation or electron transfer dissociation (Mikesh et al., 2006). It remains unknown if missing the one disulfide bridge between the β1-sheet and the DD-region could remain biologically active as the abolition of function has been demonstrated by the breakage of both (Cehovin et al., 2013).

RF2NA may fail to maintain this disulfide bridge due to constraints in the protein and DNA structures. Disulfide bonds are made by DsbA-proteins in the periplasm in these bacteria and are necessary for the biogenesis of functional type IV pili and transformability (Tinsley et al., 2004). The DD-region is of interest as it is expected to facilitate specific DUS binding (Cehovin et al., 2013) and here we made sub-divisions of the Epsilon and Gamma modes based on variable positioning of the DD-region.

**ComP-DUS modeling**

By modeling the ComP-DUS interactions, we categorize six different binding modes and one Other category based on their main interactions with either the major or minor groove of the DNA. Epsilon, Zeta and Theta were recognized by their base contacts through the DNA major groove interactions and Gamma and Delta through the minor groove. The sub-divisions of Epsilon and Gamma were based on how the DD-region differentially interacts with DNA. We interpret previous illustrated results from ComP-DUS docking experiments (Berry et al., 2016; Hughes-Games, 2020) to reflect the Epsilon mode, i.e. main interactions through the major groove.

We found strong support for the Epsilon mode across platforms and different $ComP_{nat}$. Epsilon was the most frequently modeled mode in AF3 and Chai-1 by a considerable margin, and this therefore supports Epsilon as a likely ComP-DUS complex. All platforms modeled Epsilon for three or more $ComP_{nat}$ and the Epsilon models were particularly robust and consistent in AF3 and Chai-1 for $N.sub_{nat}$, $N. muc_{nat}$, $B. den_{nat}$ and some of the $K. den_{nat}$ (Table 1; Fig 3). Fewer successful Epsilon models were found in $E. cor_{nat}$, $N. cin_{nat}$ and $N. men_{nat}$ for which more Gamma was observed.

Zeta resembles Epsilon by having the α1-β1 and β1-β2 loops entering the major groove. Zeta was not found in AF3. Zeta was always modeled together with Epsilon for the same $ComP_{nat}$ in the two platforms where Zeta was seen, Chai-1 and RF2NA, signifying the similarity of the two modes. A single Zeta model was found for N. $men_{nat}$ in Chai-1 together with several Epsilon (Table 1). In RF2NA, with less confidence (Fig 3 and 4), Zeta was predicted for K. $den_{nat}$ and E. $cor_{nat}$. We therefore consider Zeta a DD-region variant of Epsilon, recognized by an overall main interaction with bases in the major groove.

Strong support for Gamma was found in three $ComP_{nat}$, N. $cin_{nat}$, N. $men_{nat}$ and E. $cor_{nat}$. Together all platforms modeled Gamma for at least one of these three $ComP_{nat}$. A few successful Gamma models were also observed in the remaining four N.$sub_{nat}$, B. $den_{nat}$, K.$den_{nat}$ and N. $muc_{nat}$ and then only by RF2NA/HADDOCK. Although less confidence can be attributed RF2NA models overall (Figure 3 and 4), it is notable that successful Gamma is predicted at least once for all $ComP_{nat}$. Gamma is very different from Epsilon/Zeta in having the α 1- β 1, β 1- β 2 loops entering the minor groove.

Like Zeta can be considered a variant of Epsilon, Delta resembles Gamma in having its main interactions with the minor groove. When Delta was found for any $ComP_{nat}$, the same platform also modeled Gamma showing a correlation between the two modes as with Epsilon/Zeta. These mode and sub-mode models may reflect a flexibility of the DD-region which may be functionally relevant. Delta was modeled for N. $sub_{nat,}$ N. $cin_{nat,}$ N. $men_{nat}$ and E. $cor_{nat}$ by RF2NA and below ipTM cutoff in AF3 (N. $men_{nat}$). N. $sub_{nat}$ was the only modeled complex in this RF2NA Delta subset with also very robust Epsilon models in AF3 and Chai-1. In modeling predictions yielding Delta, N. $cin_{nat,}$ N. $men_{nat}$ and E. $cor_{nat}$ stand out as complexes imposing modeling challenges for all platforms, although the latter achieves moderately high

ipTM scores in AF3 and Chai-1. E. $cor_{nat}$ was also the only complex with a substantial fraction of highly divergent modes assigned Other (Fig. 2-3; Table 1).

The Theta mode was uniquely identified in N.$cin_{nat}$ and only in Chai-1. Unlike all other modes, Theta showed major conformational changes in the DNA, complete strand separation, which potentially could allow better access to bases. Theta resembles Epsilon/Zeta in that the major interactions were in the major groove, which was pried apart in Theta. The consistently flipped out base is the central thymine in the conserved 5'-CTG-3' DUS motif alluding to essential involvement in DUS specific binding. The only other severe alteration of the DNA structure was found in N. $sub_{nat}$ Epsilon Chai-1 models, also displaying a flipped out base. This fifth thymine base immediately precedes the conserved 5'-CTG-3' core of the AG-DUS and it is tempting to speculate that the modeled conformational change in the DNA may instigate specific binding. Further experimentation is required to investigate these models in depth by manipulation of input parameters and MSA schemes.

The Eta model was never predicted by AF3, Chai-1 or RF2NA, but appeared in the re-docking experiments using HADDOCK. Eta was therefore not considered a supported mode, but draws attention to the ability of re-docking to both reproduce modeled complexes and predict new. HADDOCK did robustly generate re-docks showing Epsilon and Gamma modes in support of these (Table 1).

Other modes than the above were considered anomalies in modeling E.$cor_{nat}$ in Chai-1 and in a single model in B. $den_{nat}$ in RF2NA. As visualized in Fig. 2 for E.$cor_{nat}$, these alternative modes are characterized by binding to the opposite end of ComP relative to all other models. If anything, these findings again show the challenge in robustly modeling E.$cor_{nat}$ and that the RF2NA outputs are more variable and of lower quality than those of AF3 and Chai-1.

Generally, the RF2NA predicted less confident models than AF3 and Chai-1. This observation aligns with those of Abramson et al. (2024) demonstrating RF2NA to be inferior to AF3 for protein-nucleic acid prediction on benchmark sets of the PDB database, having about 40% lower success rate than AF3.

The generally high PAE values in successful RF2NA models observed herein contrasts the unity of nearly identical AF3 models, all with much lower PAE values relative to the RF2NA models, speaking for the robustness of the AF3 models. Although there is some variation within successful Chai-1 models, the consistent overlap with AF3 in terms of Epsilon and Gamma is striking. Higher variation in the Chai-1 models compared to those from AF3 may also be explained by Chai-1's variable MSA input, whereas the AF3 web server does not allow for MSA customization. Thus, the cross-platform deviance likely stems from the AF3 and Chai-1 results being of generally better quality than the those from RF2NA. However, all platforms may prove informative when exploring DUS-specificity. It is also paradoxical that even in the cases where the models are of putatively high quality in both programs (RF2NA: PAE < 10; AF3: ipTM $\geq$ 0.8), the modes are usually Epsilon or Gamma. However, we cannot exclude that there might be ComP-DUS complexes that RF2NA is able to solve while AF3 is not. For instance, RF2NA may have yielded insight into *N. meningitidis* MC58$_{\text{ComP}}$-AT-DUS and *N. cinerea* ATCC 14685$_{\text{ComP}}$-AT-DUS, as both of these had models of PAE <10, while AF3 always failed to predict these combinations. At the same time, we never observed any RF2NA models of higher quality than any AF3 or Chai-1 models in terms of PAE, which is striking.

The seeming challenges RF2NA has with homing in on specific ComP-DUS complexes may not necessarily extend to other proteins-NA complexes. Although it is unclear to which degree RF2NA can return “false positive” models of high confidence that fail to recover the true binding mode, Baek et al. (2024) reported that 81% of models with PAE <10 was correctly predicted the interface of the complex, which is fairly high.

In exploring the AF3 modeling quality of DUS against scrambled DUS (Fig. 5), $ComP_{N.cin}$ and $ComP_{N.men}$ were found as the only ComPs where median scrambled DUS models were significantly better than native pairings. Experimental data by Cehovin et al. (2013) showed that, in addition to DUS-specific transformation, non-DUS transformation is also mediated by $ComP_{N.men}$. These *in vivo* results showed not only the well-characterized wild-type bias for DUS-specific transformation, but also that non-DUS transformation could reach the level of DUS-specific transformation when overexpressing ComP (Cehovin et al., 2013). We propose that this DUS/non-DUS ambiguity could be due to two distinct binding modes identified here, Epsilon (Zeta/Theta) and Gamma (Delta). However, since $ComP_{N.\ men}$ was inconsistently predicted to belong to any of these different modes in all platforms, including HADDOCK, we cannot herein find strong support for either mode being the DUS-specific mode, and alternative tests should be devised to test this hypothesis experimentally in e.g. the *N. meningitidis* or *N. subflava* model organisms.

**Figure 5.** **A:** Violin plots showing distributions of ipTM for native ("Native") and scrambled DUS ("Scrambled") runs in AF3 for the docking set. Box plots encompassing the median and interquartile ranges are seen inside the violins. AA-king3DUS (K. den). AG-DUS (B. den + N. sub). AG-eikDUS (E. cor). AG-mucDUS (N. muc). AT-DUS (N. men + N. cin). **B:** Same as in A for CPPM. Refer to Code Output S2 for statistical significance tests for Native vs. Scrambled for all DUS dialects.

Further assessing AF3 modeling quality with the scrambled DUS control, we found that $ComP_{K.den}$ did not significantly differentiate between its native DUS AA-king3-DUS and the scrambled versions. $K.den_{nat}$ in AF3 were consistently and robustly modeled in Epsilon in all three platforms and in HADDOCK. Based on this finding we cannot exclude the possibility

that Epsilon represents a non-specific mode, if there is one for $ComP_{K.den}$. In contrast to the robust Epsilon modes, only four and not strong Gamma models were modeled by RF2NA providing weak support for Gamma as the DUS-specific mode for $ComP_{K.den}$.

In contrast to the above, $ComP_{B.\ den}$/$ComP_{N.\ sub}$, $ComP_{N.\ muc}$ and $ComP_{E.\ cor}$ models were significantly better for native complexes than scrambled. These results suggest that AF3 more accurately predicts the DUS-specific mode than any non-specific mode in these ComPs. Assigning either Epsilon or Gamma as the one DUS-specific mode remains unresolved as also these $ComP_{nat}$ were confidently modeled to either Epsilon (B. $den_{nat}$, N. $sub_{nat}$ and N. $muc_{nat}$) and Gamma (E. $cor_{nat}$). Together these results give direction for further studies.

# Conclusions

We find for the first time strong support for ComP to bind DUS through two distinct modes, Epsilon and Gamma, applying AF3, Chai-1 and RF2NA modeling platforms. These modes differ in their respective interactions with either the DNA major or minor groove. Strong modeling support for Epsilon was found in N. $sub_{nat}$, B. $den_{nat}$, K. $den_{nat}$ and N. $muc_{nat}$, and Gamma mainly found its support in E. $cor_{nat}$, N. $cin_{nat}$ and N. $men_{nat}$. Both Epsilon and Gamma found support in re-docking experiments using HADDOCK. It remains to be investigated if one of the modes represents DUS-specific binding and the other non-specific binding. Differences in amino acid composition of ComPs, as well as minor differences in structural domains in ComPs such as the predicted α-turns in AF3 $ComP_{B.\ den}$, $ComP_{E.\ cor}$, and $ComP_{N.\ muc}$ identified here, might help understand specific binding to distinct DUS-dialects.

# Data availability

Scripts and data used are provided in the GitHub repository https://github.com/stianale/ComP-DUS-modeling.

# Author contributions

SH: Data curation, formal analysis, investigation, methodology, implementation of computer codes and supporting algorithms, data presentation, writing original draft and review & editing. AHL; KA and SAF: Supervision of SH and scientific critical review, writing – review &

editing. OHA: Conceptualization, funding acquisition, project administration, resources, supervision, visualization, writing – review & editing

# References


Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., … Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, *630*(8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w

Ambur, O. H., Engelstädter, J., Johnsen, P. J., Miller, E. L., & Rozen, D. E. (2016). Steady at the wheel: Conservative sex and the benefits of bacterial transformation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1706). https://doi.org/10.1098/rstb.2015.0528

Baek, M., McHugh, R., Anishchenko, I., Jiang, H., Baker, D., & DiMaio, F. (2024). Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nature Methods*, *21*(1), 117–121. https://doi.org/10.1038/s41592-023-02086-5

Basu, S., & Wallner, B. (2016). DockQ: A Quality Measure for Protein-Protein Docking Models. *PLOS ONE*, *11*(8), e0161879. https://doi.org/10.1371/journal.pone.0161879

Berry, J.-L., Cehovin, A., McDowell, M. A., Lea, S. M., & Pelicic, V. (2013). Functional Analysis of the Interdependence between DNA Uptake Sequence and Its Cognate

ComP Receptor during Natural Transformation in Neisseria Species. *PLoS Genetics*, *9*(12), e1004014. https://doi.org/10.1371/journal.pgen.1004014

Berry, J.-L., Xu, Y., Ward, P. N., Lea, S. M., Matthews, S. J., & Pelicic, V. (2016). A Comparative Structure/Function Analysis of Two Type IV Pilin DNA Receptors Defines a Novel Mode of DNA Binding. *Structure*, *24*(6), 926–934. https://doi.org/10.1016/j.str.2016.04.001

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Cehovin, A., Simpson, P. J., McDowell, M. A., Brown, D. R., Noschese, R., Pallett, M., Brady, J., Baldwin, G. S., Lea, S. M., Matthews, S. J., & Pelicic, V. (2013). Specific DNA recognition mediated by a type IV pilin. *Proceedings of the National Academy of Sciences*, *110*(8), 3065–3070. https://doi.org/10.1073/pnas.1218832110

Chai Discovery. (2024). Chai-1: Decoding the molecular interactions of life. *bioRxiv*. https://doi.org/10.1101/2024.10.10.615955

Corso, G., Stärk, H., Jing, B., Barzilay, R., & Jaakkola, T. (2022). *DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking*. https://doi.org/10.48550/arxiv.2210.01776

Dubnau, D., & Blokesch, M. (2019). Mechanisms of DNA Uptake by Naturally Competent Bacteria. *Annual Review of Genetics*, *53*, 217–237. https://doi.org/10.1146/annurev-genet-112618-043641

Duffin, P. M., & Seifert, H. S. (2012). Genetic transformation of Neisseria gonorrhoeae shows a strand preference. *FEMS Microbiology Letters*, *334*(1), 44–48. https://doi.org/10.1111/j.1574-6968.2012.02612.x

Esmaeeli, R., Bauzá, A., & Perez, A. (2023). Structural predictions of protein–DNA binding: MELD-DNA. *Nucleic Acids Research*, *51*(4), 1625–1636. https://doi.org/10.1093/nar/gkad013

Frye, S. A., Nilsen, M., Tønjum, T., & Ambur, O. H. (2013). Dialects of the DNA Uptake Sequence in Neisseriaceae. *PLOS Genetics*, *9*(4), e1003458. https://doi.org/10.1371/journal.pgen.1003458

Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A., & Caves, L. S. D. (2006). Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics*, *22*(21), 2695–2696. https://doi.org/10.1093/bioinformatics/btl461

Grant, B. J., Skjærven, L., & Yao, X. (2021). The BIO3D packages for structural bioinformatics. *Protein Science*, *30*(1), 20–30. https://doi.org/10.1002/pro.3923

Hughes-Games, A. (2020). *New Approaches to Characterising Mechanisms of Horizontal Gene Transfer in Neisseria Gonorrhoeae Using Microscopy and Genome Sequence Analysis*. ProQuest Dissertations & Theses.

João M. C. Teixeira, Rodrigo Vargas Honorato, Marco Giulini, Alexandre Bonvin, SarahAlidoost, Victor Reys, Brian Jimenez, Douwe Schulte, Charlotte van Noort, Stefan Verhoeven, Barbara Vreede, SSchott, & Regen Tsai. (2024). *haddocking/haddock3: V3.0.0-beta.5* (Version v3.0.0-beta.5) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.10527751

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G. R., Morey-Burrows, F. S., Anishchenko, I., Humphreys, I. R., McHugh, R., Vafeados, D., Li, X., Sutherland, G. A., Hitchcock, A., Hunter, C. N., Kang, A., Brackenbrough, E., Bera, A. K., … Baker, D. (n.d.). Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, *384*(6693), eadl2528. https://doi.org/10.1126/science.adl2528

Mikesh, L. M., Ueberheide, B., Chi, A., Coon, J. J., Syka, J. E. P., Shabanowitz, J., & Hunt, D. F. (2006). The utility of ETD mass spectrometry in proteomic analysis. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1764*(12), 1811–1822. https://doi.org/10.1016/j.bbapap.2006.10.003

Mirabello, C., & Wallner, B. (2024). *DockQ v2: Improved automatic quality measure for protein multimers, nucleic acids, and small molecules*. Bioinformatics. https://doi.org/10.1101/2024.05.28.596225

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: Making protein folding accessible to all. *Nature Methods*, *19*(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1

Qiao, Z., Nie, W., Vahdat, A., Miller, T. F., & Anandkumar, A. (2024). State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence*, *6*(2), 195–208. https://doi.org/10.1038/s42256-024-00792-z

Skjærven, L., Jariwala, S., Yao, X.-Q., & Grant, B. J. (2016). Online interactive analysis of protein structure ensembles with Bio3D-web. *Bioinformatics*, *32*(22), 3510–3512. https://doi.org/10.1093/bioinformatics/btw482

Skjærven, L., Yao, X.-Q., Scarabelli, G., & Grant, B. J. (2014). Integrating protein structural dynamics and evolutionary analysis with Bio3D. *BMC Bioinformatics*, *15*(1), 399. https://doi.org/10.1186/s12859-014-0399-6

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., & Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, *20*(1), 473. https://doi.org/10.1186/s12859-019-3019-7

Steinegger, M., Mirdita, M., & Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods*, *16*(7), 603–606. https://doi.org/10.1038/s41592-019-0437-4

Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications*, *9*(1), 2542. https://doi.org/10.1038/s41467-018-04964-5

Team, P. (2024). Technical Report of HelixFold3 for Biomolecular Structure Prediction. *arXiv*. https://doi.org/10.48550/arXiv.2408.16975

Tinsley, C. R., Voulhoux, R., Beretti, J.-L., Tommassen, J., & Nassif, X. (2004). Three homologues, including two membrane-bound proteins, of the disulfide oxidoreductase DsbA in Neisseria meningitidis: Effects on bacterial growth and biogenesis of functional type IV pili. *J Biol Chem*, *279*(26), 27078–27087. https://doi.org/10.1074/jbc.M313404200

Treangen, T. J., Ambur, O. H., Tonjum, T., & Rocha, E. P. C. (2008). The impact of the neisserial DNA uptake sequences on genome evolution and stability. *Genome Biology*, *9*(3), R60. https://doi.org/10.1186/gb-2008-9-3-r60

Wohlwend, J., Corso, G., Passaro, S., Reveiz, M., Leidal, K., Swiderski, W., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T., & Barzilay, R. (2024). Boltz-1: Democratizing Biomolecular Interaction Modeling. bioRxiv. https://doi.org/10.1101/2024.11.19.624167

Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, *57*(4), 702–710. https://doi.org/10.1002/prot.20264

## Supporting information captions

**Table S1** ComPs used for the different species and the paired DUS-dialect.

**Table S2** The different MSA schemes used in Chai-1, listing the various settings.

**Figure S1.** Representative ComP-DUS structures showing the various binding modes found in AF3, Chai-1 and RF2NA.

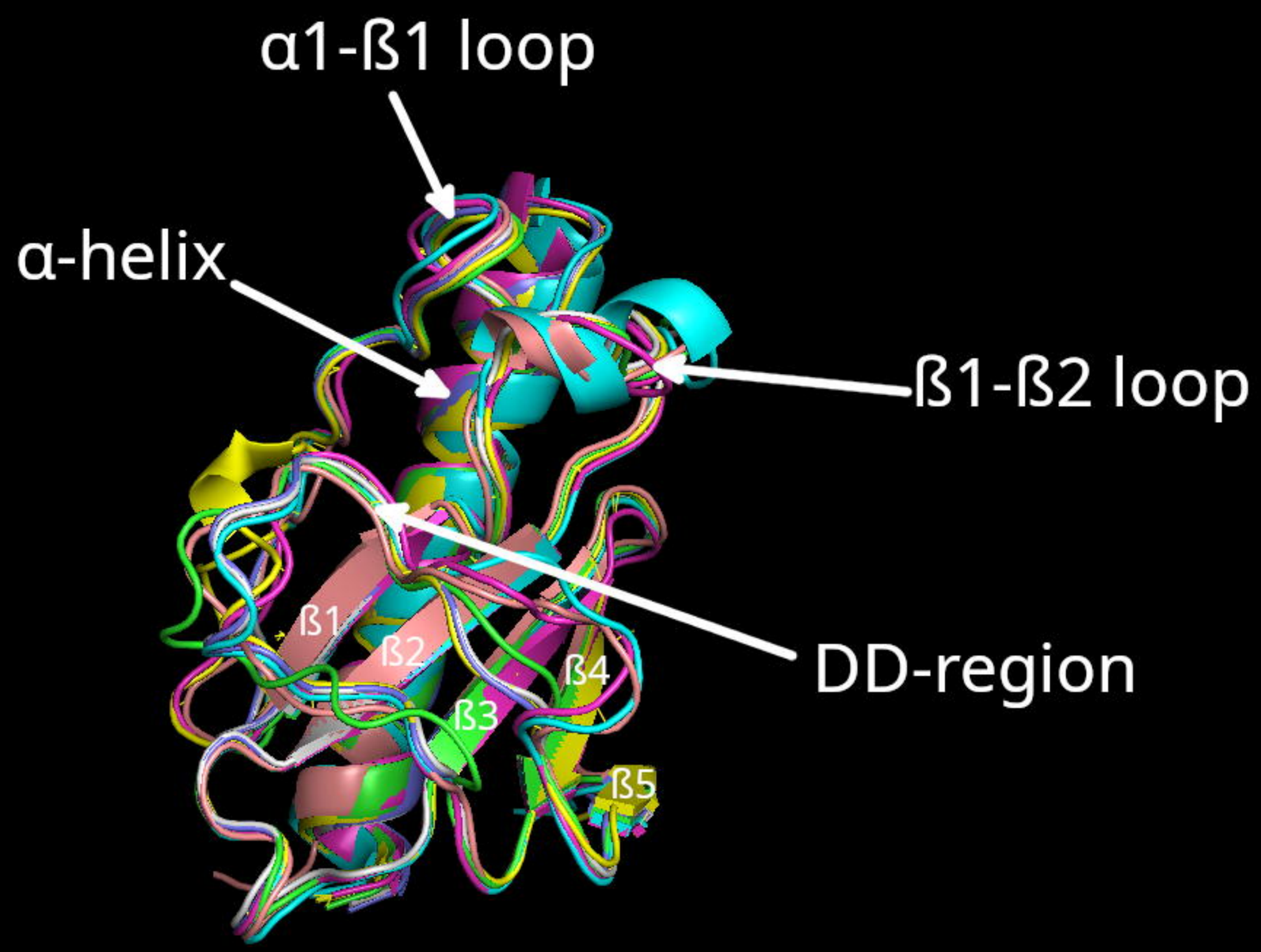
**Figure S2.** Alignment of the ComPs, using default settings in MAFFT v. 7.505, visualized in MVIEW v. 1.68. First 28 amino acids of the N. meningitidis mature ComP are trimmed and the other ComPs accordingly.

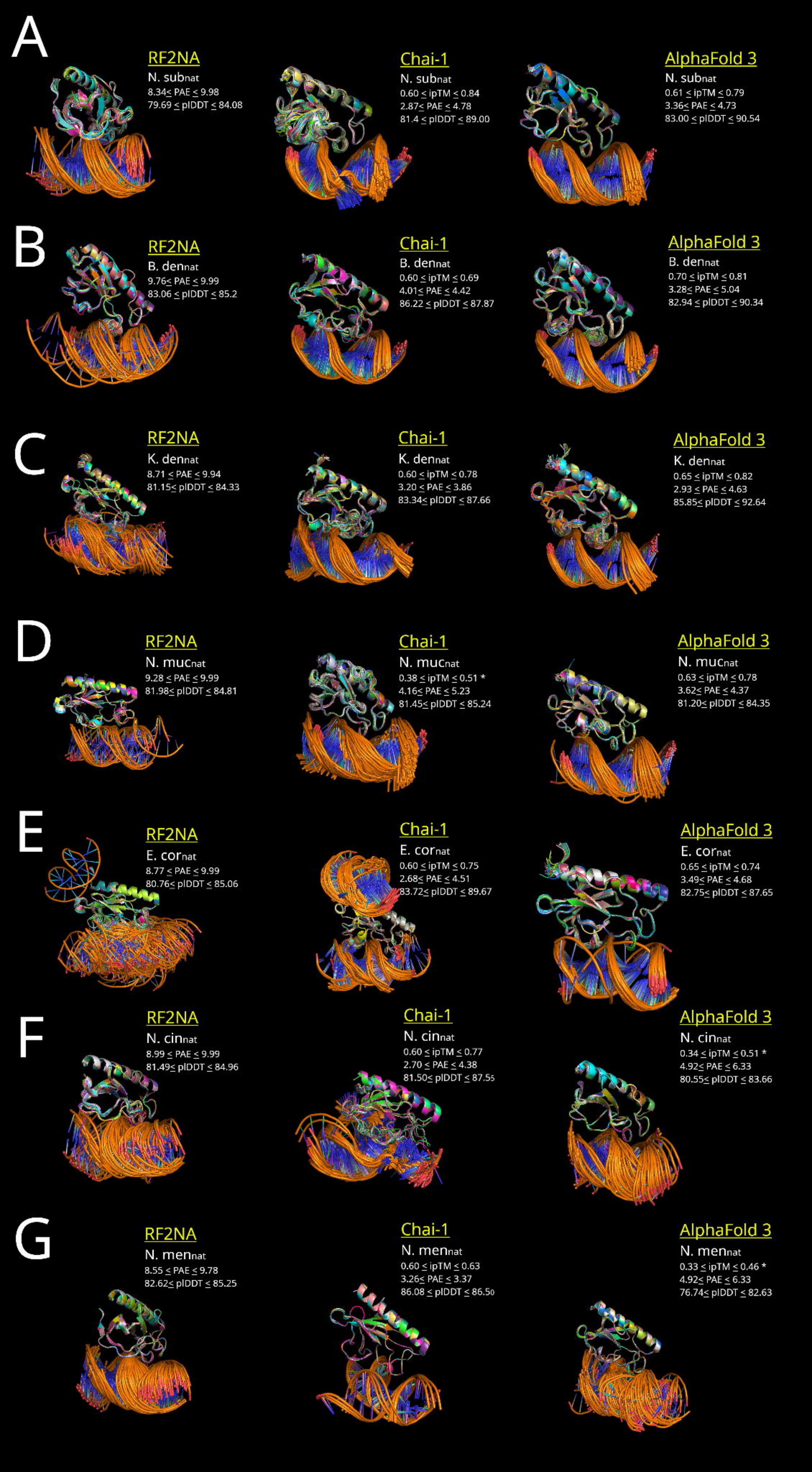
**Figure S3.** Selected high-ranking HADDOCK models, representative of the dominating binding mode. A: N. $muc_{nat}$ (Epsilon). B: N. $men_{nat}$ (Epsilon). C: K. $den_{nat}$ (Epsilon). D: E. $cor_{nat}$ (Epsilon). E: B. $den_{nat}$ (Epsilon). F: N. $cin_{nat}$ (Gamma). G: N. $sub_{nat}$ (Gamma).

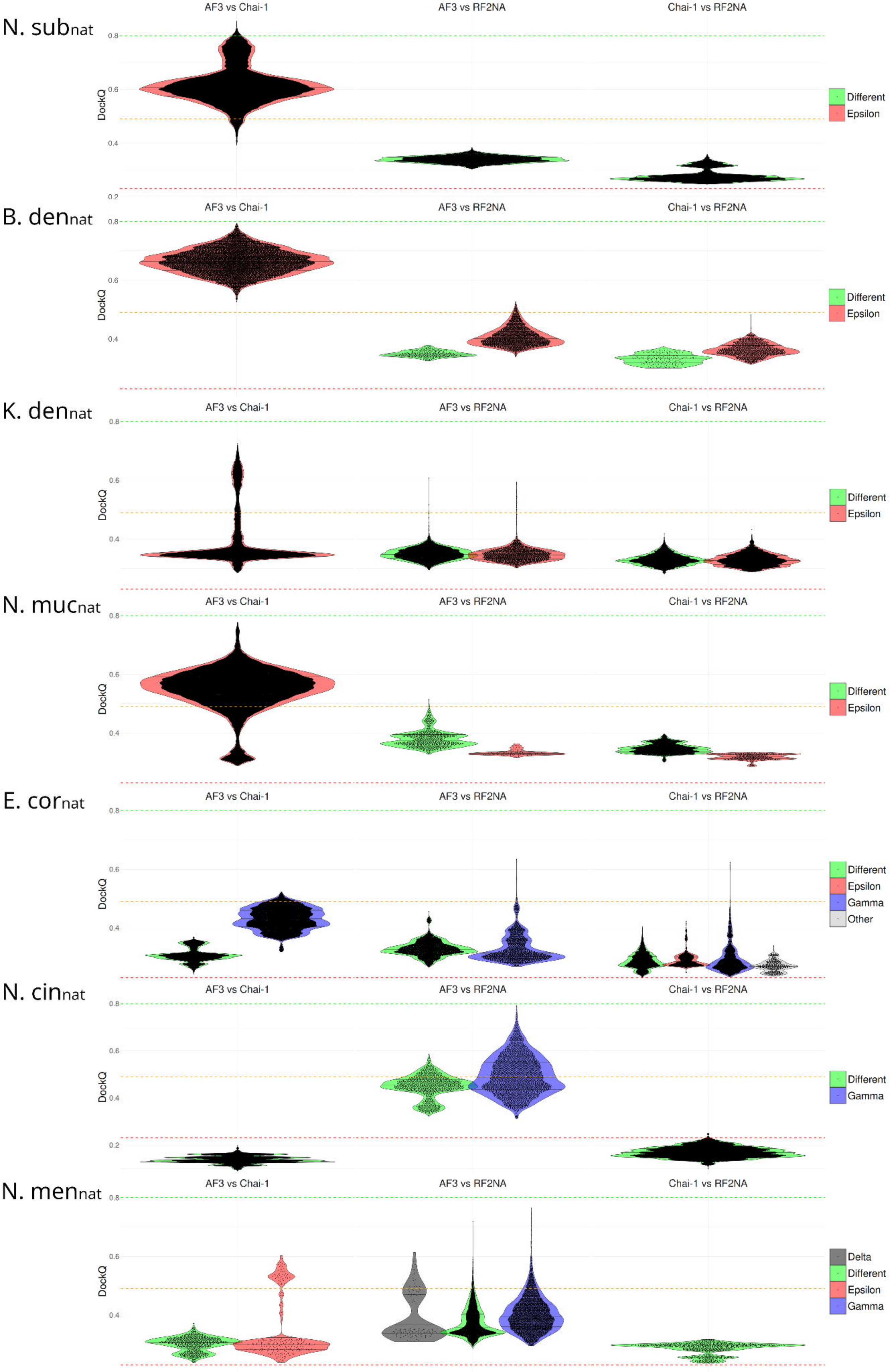
**Code Output S1** Wilcoxon rank-sum tests on PAE and pLDDT, comparing AF3, Chai-1 and RF2NA. Median PAE and pLDDT values are also shown.

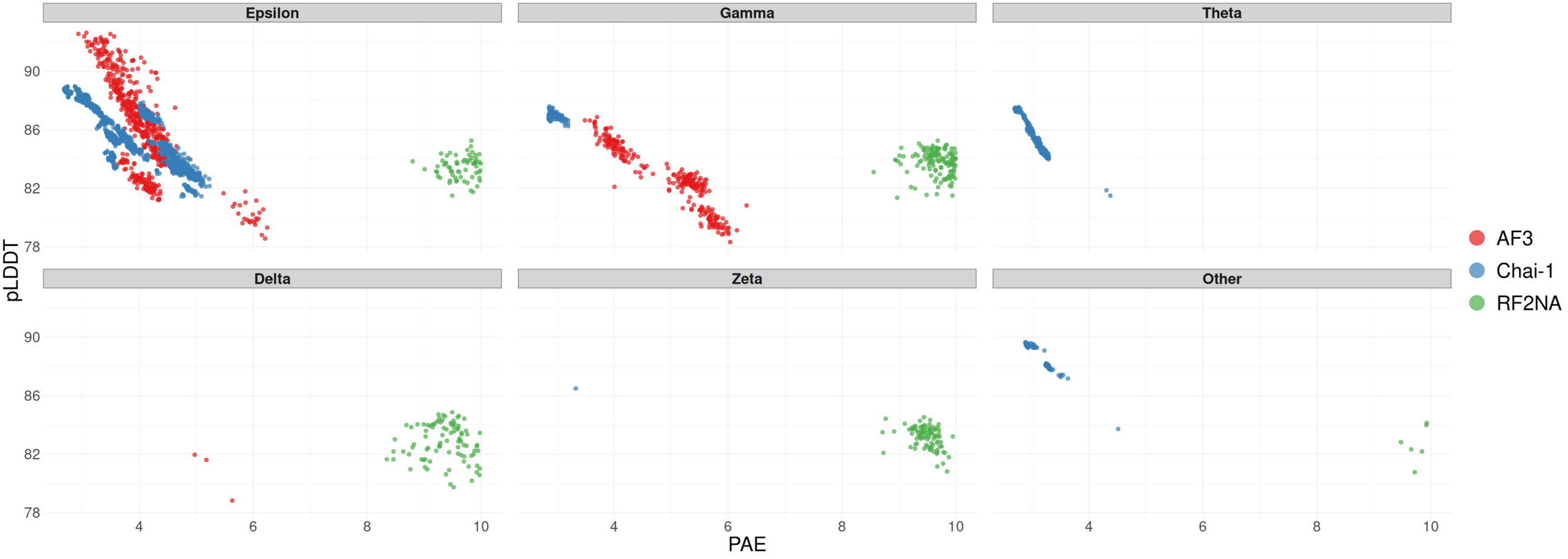
**Code Output S2** Wilcoxon rank-sum tests on ipTM and CPPM for each $ComP_{nat}$.

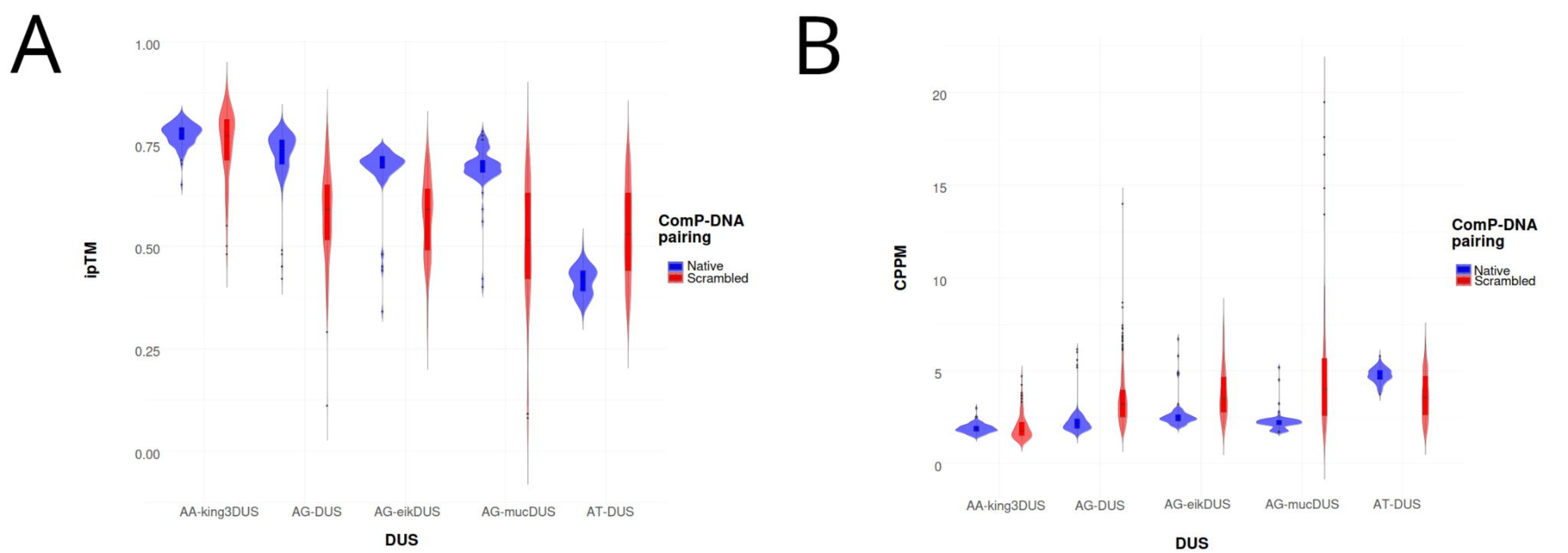
**Code Output S3** Wilcoxon rank-sum test on DockQ scores for the internal platform consistency check.


α1-ß1 loop
α-helix
ß1-ß2 loop
DD-region
ß1
ß2
ß3
ß4
ß5


A
RF2NA
N. subnat
8.34≤ PAE ≤ 9.98
79.69 ≤ plDDT ≤ 84.08
Chai-1
N. subnat
0.60 ≤ ipTM ≤ 0.84
2.87≤ PAE ≤ 4.78
81.4 ≤ plDDT ≤ 89.00
AlphaFold 3
N. subnat
0.61 ≤ ipTM ≤ 0.79
3.36≤ PAE ≤ 4.73
83.00 ≤ plDDT ≤ 90.54
B
RF2NA
B. dennat
9.76≤ PAE ≤ 9.99
83.06 ≤ plDDT ≤ 85.2
Chai-1
B. dennat
0.60 ≤ ipTM ≤ 0.69
4.01≤ PAE ≤ 4.42
86.22 ≤ plDDT ≤ 87.87
AlphaFold 3
B. dennat
0.70 ≤ ipTM ≤ 0.81
3.28≤ PAE ≤ 5.04
82.94 ≤ plDDT ≤ 90.34
C
RF2NA
K. dennat
8.71 ≤ PAE ≤ 9.94
81.15≤ plDDT ≤ 84.33
Chai-1
K. dennat
0.60 ≤ ipTM ≤ 0.78
3.20 ≤ PAE ≤ 3.86
83.34≤ plDDT ≤ 87.66
AlphaFold 3
K. dennat
0.65 ≤ ipTM ≤ 0.82
2.93 ≤ PAE ≤ 4.63
85.85≤ plDDT ≤ 92.64
D
RF2NA
N. mucnat
9.28 ≤ PAE ≤ 9.99
81.98≤ plDDT ≤ 84.81
Chai-1
N. mucnat
0.38 ≤ ipTM ≤ 0.51 *
4.16≤ PAE ≤ 5.23
81.45≤ plDDT ≤ 85.24
AlphaFold 3
N. mucnat
0.63 ≤ ipTM ≤ 0.78
3.62≤ PAE ≤ 4.37
81.20≤ plDDT ≤ 84.35
E
RF2NA
E. cornat
8.77 ≤ PAE ≤ 9.99
80.76≤ plDDT ≤ 85.06
Chai-1
E. cornat
0.60 ≤ ipTM ≤ 0.75
2.68≤ PAE ≤ 4.51
83.72≤ plDDT ≤ 89.67
AlphaFold 3
E. cornat
0.65 ≤ ipTM ≤ 0.74
3.49≤ PAE ≤ 4.68
82.75≤ plDDT ≤ 87.65
F
RF2NA
N. cinnat
8.99 ≤ PAE ≤ 9.99
81.49≤ plDDT ≤ 84.96
Chai-1
N. cinnat
0.60 ≤ ipTM ≤ 0.77
2.70 ≤ PAE ≤ 4.38
81.50≤ plDDT ≤ 87.55
AlphaFold 3
N. cinnat
0.34 ≤ ipTM ≤ 0.51 *
4.92≤ PAE ≤ 6.33
80.55≤ plDDT ≤ 83.66
G
RF2NA
N. mennat
8.55 ≤ PAE ≤ 9.78
82.62≤ plDDT ≤ 85.25
Chai-1
N. mennat
0.60 ≤ ipTM ≤ 0.63
3.26≤ PAE ≤ 3.37
86.08 ≤ plDDT ≤ 86.50
AlphaFold 3
N. mennat
0.33 ≤ ipTM ≤ 0.46 *
4.92≤ PAE ≤ 6.33
76.74≤ plDDT ≤ 82.63


N. subnat
B. dennat
K. dennat
N. mucnat
E. cornat
N. cinnat
N. mennat
AF3 vs Chai-1
AF3 vs RF2NA
Chai-1 vs RF2NA
DockQ
0.8
0.6
0.4
0.2
Different
Epsilon
Gamma
Other
Delta


Epsilon
Gamma
Theta
Delta
Zeta
Other
pLDDT
PAE
90
86
82
78
4
6
8
10
AF3
Chai-1
RF2NA


A
1.00
0.75
0.50
0.25
0.00
ipTM
AA-king3DUS
AG-DUS
AG-eikDUS
AG-mucDUS
AT-DUS
DUS
ComP-DNA pairing
Native
Scrambled
B
20
15
10
5
0
CPPM
AA-king3DUS
AG-DUS
AG-eikDUS
AG-mucDUS
AT-DUS
DUS
ComP-DNA pairing
Native
Scrambled