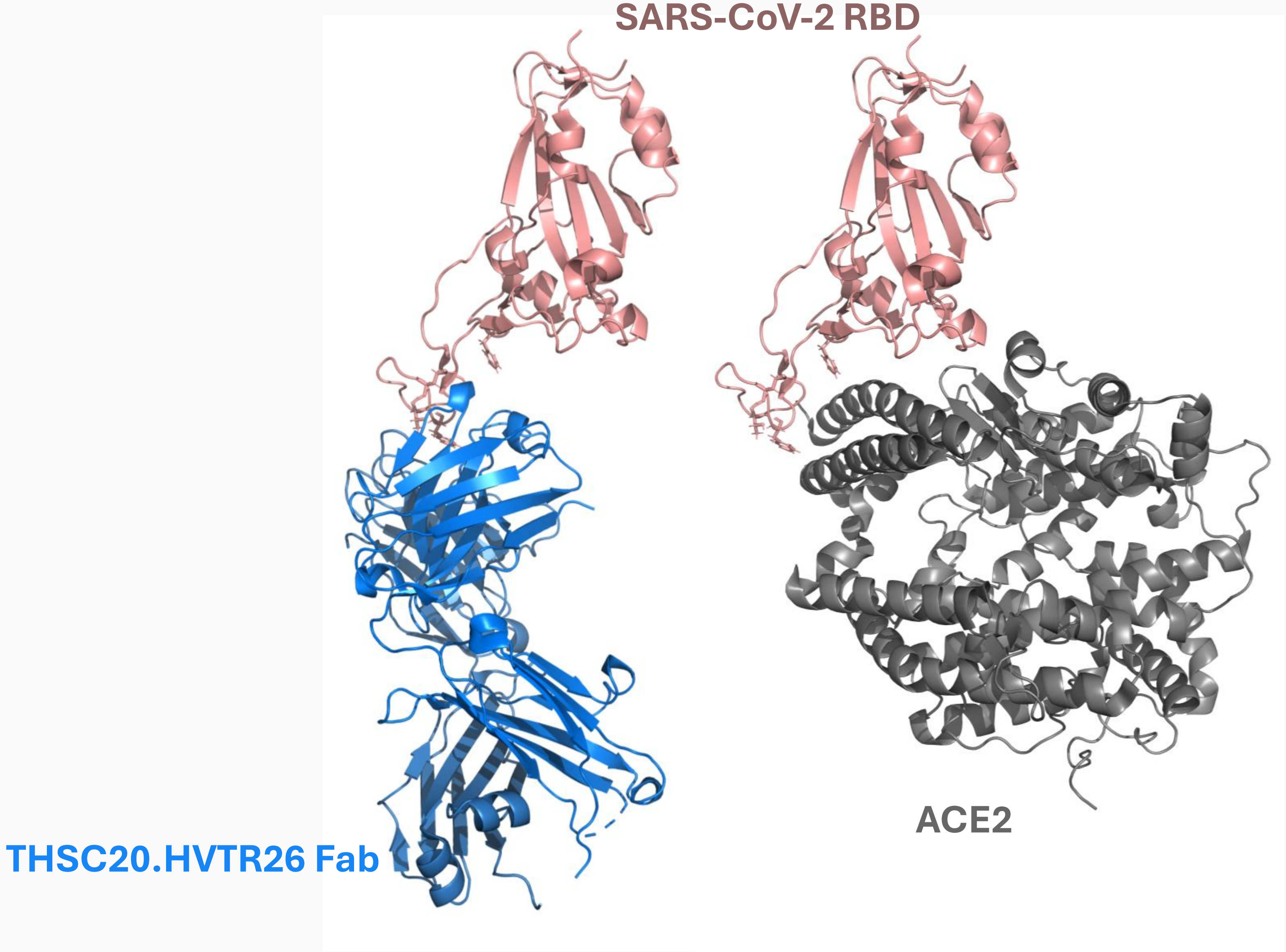
ProteinMPNN-based binding interface analytical pipeline

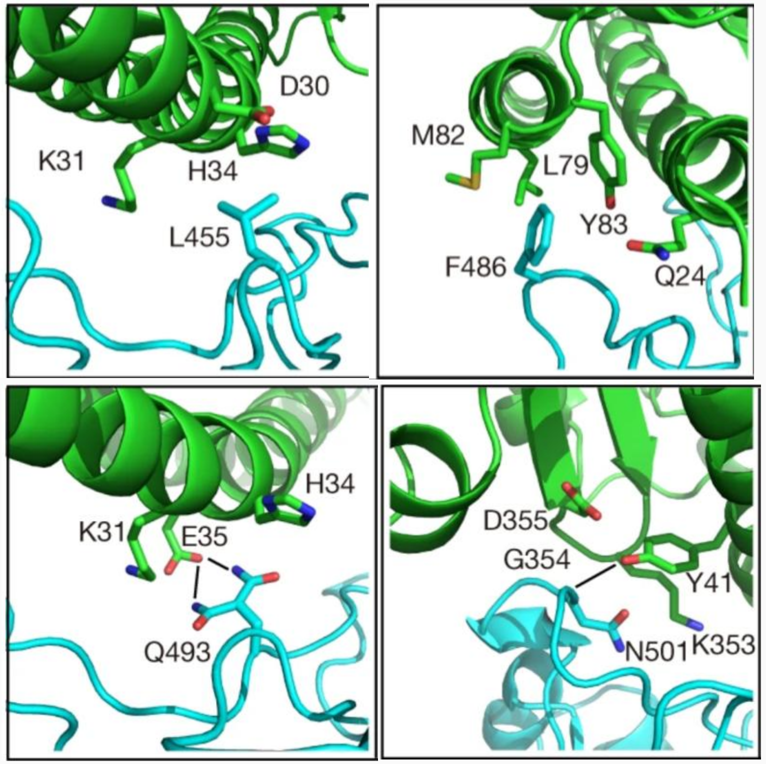
Tiejun Wei
Research Fellow
Jone’s Lab,
Department of Computer Science,
University College London

Quick structure view



S1–ACE complex

- Networks of hydrophilic interactions
  - 13 Hydrogen bonds 2 salt bridges at the RBD side interface.
- Multiple Tyr –> H–bond with the polar hydroxyl group.
  - [T449,T489,T505]
- Key contact residues shown as righ figure

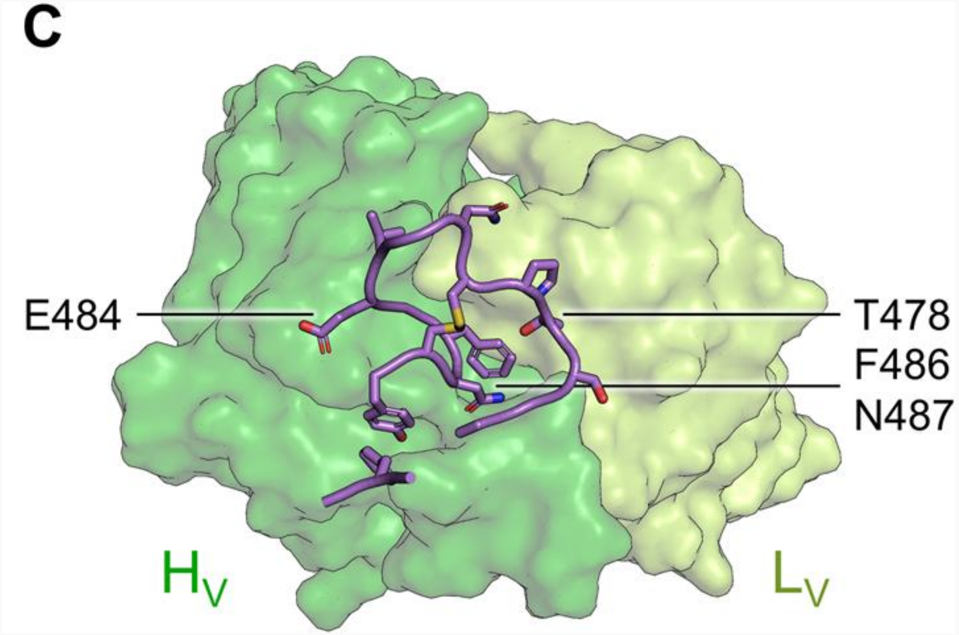


https://www.nature.com/articles/s41586-020-2180-5

S1-Ab complex

- A few key contact residues on Spike protein has been identified.
- Overlaping residues:

  [478, 484, 486, 487]



https://pubmed.ncbi.nlm.nih.gov/35482816/

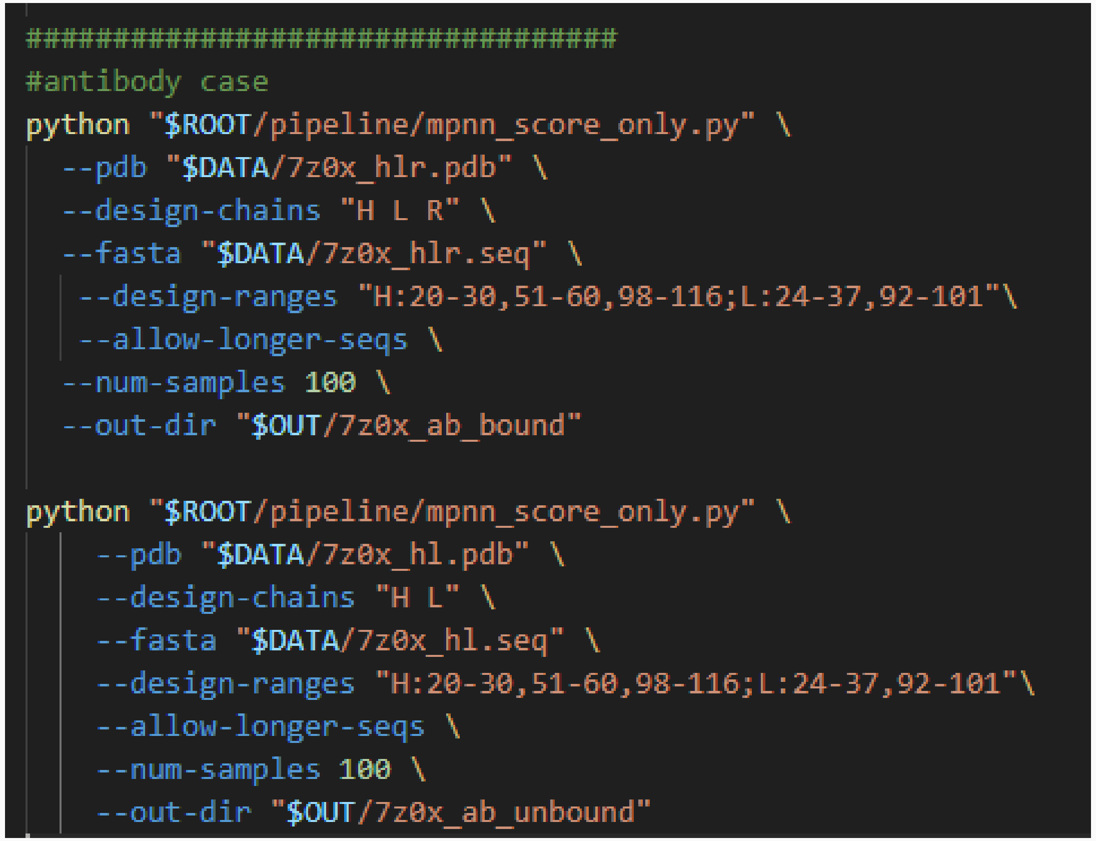
Reframed problem:

- Given a known structure, we inverse-fold the paired complex:
  - Observable:
  - $p(\mathrm{Seq}_A \mid \mathrm{Structure}_{AB})$ and $p(\mathrm{Seq}_A \mid \mathrm{Structure}_A)$
  - Hypothesis:
  - Interface hotspot correlates to:
    - $\Delta\mathrm{NLL} = \mathrm{NLL}_{\mathrm{bound}} - \mathrm{NLL}_{\mathrm{unbound}}$
    - Entropy $H_i = -\sum p_i(a) \log pi(a)$
    - Mutation penalties: $\mathrm{loss}_a = \log p_i(\mathrm{WT}) - \log p_i(a)$
  - We measure ‘bound’ structure constraint added on top of the sequence.

Pipeline overview

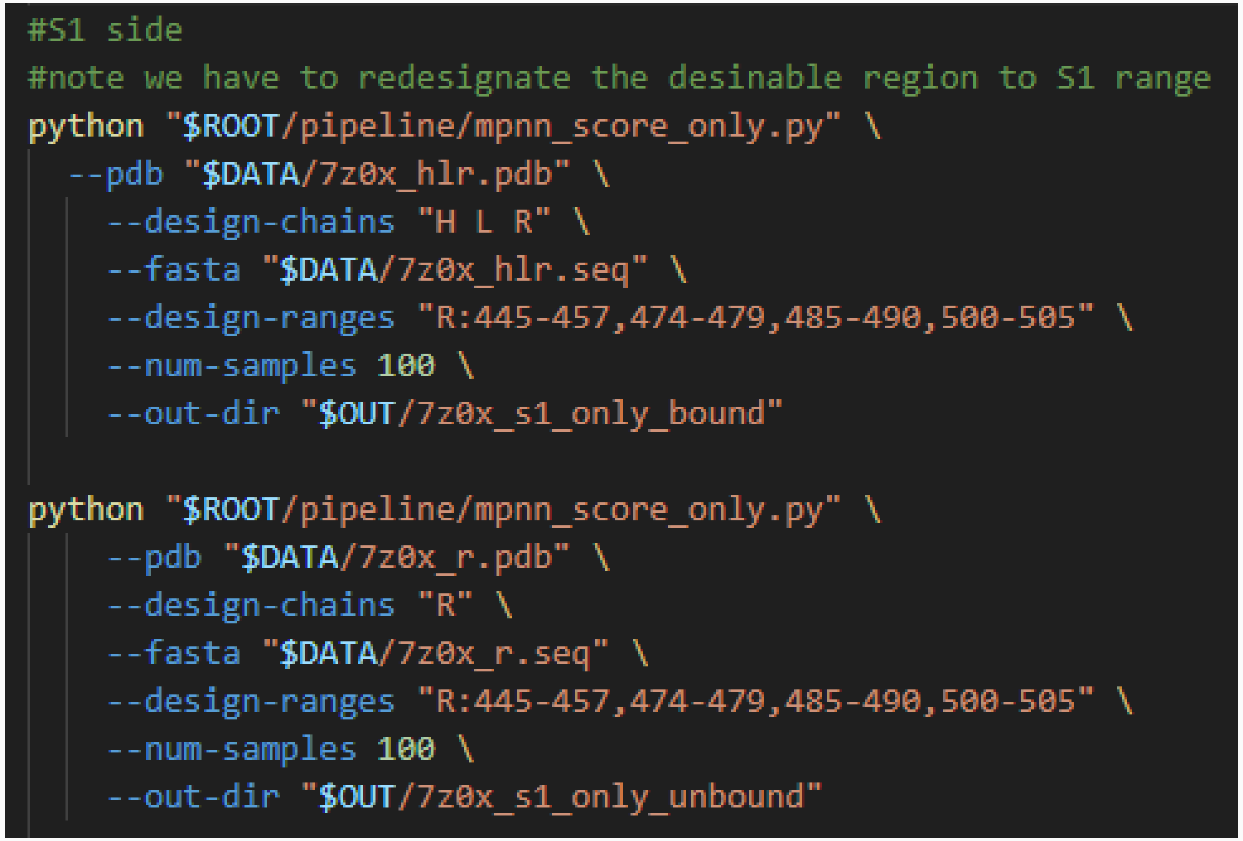
- 1. preliminary analysis:
  - Contact map
  - H-bond count
  - Shared residue sets between pair A-B/A-C
  - Biophysical calculations (metaD, FEP etc)
- 2. inverse fold confidence evaluation
- 3. cross-ref/ bagging amongst methods,.
- 4. summarize/plot

Example script:



```
##################################
#antibody case
python "$ROOT/pipeline/mpnn_score_only.py" \
  --pdb "$DATA/7z0x_hlr.pdb" \
  --design-chains "H L R" \
  --fasta "$DATA/7z0x_hlr.seq" \
   --design-ranges "H:20-30,51-60,98-116;L:24-37,92-101"\
   --allow-longer-seqs \
  --num-samples 100 \
  --out-dir "$OUT/7z0x_ab_bound"

python "$ROOT/pipeline/mpnn_score_only.py" \
    --pdb "$DATA/7z0x_hl.pdb" \
    --design-chains "H L" \
    --fasta "$DATA/7z0x_hl.seq" \
    --design-ranges "H:20-30,51-60,98-116;L:24-37,92-101"\
    --allow-longer-seqs \
    --num-samples 100 \
    --out-dir "$OUT/7z0x_ab_unbound"
```



```
#S1 side
#note we have to redesignate the desinable region to S1 range
python "$ROOT/pipeline/mpnn_score_only.py" \
  --pdb "$DATA/7z0x_hlr.pdb" \
    --design-chains "H L R" \
    --fasta "$DATA/7z0x_hlr.seq" \
    --design-ranges "R:445-457,474-479,485-490,500-505" \
    --num-samples 100 \
    --out-dir "$OUT/7z0x_s1_only_bound"

python "$ROOT/pipeline/mpnn_score_only.py" \
    --pdb "$DATA/7z0x_r.pdb" \
    --design-chains "R" \
    --fasta "$DATA/7z0x_r.seq" \
    --design-ranges "R:445-457,474-479,485-490,500-505" \
    --num-samples 100 \
    --out-dir "$OUT/7z0x_s1_only_unbound"
```

What does ΔNLL mean?

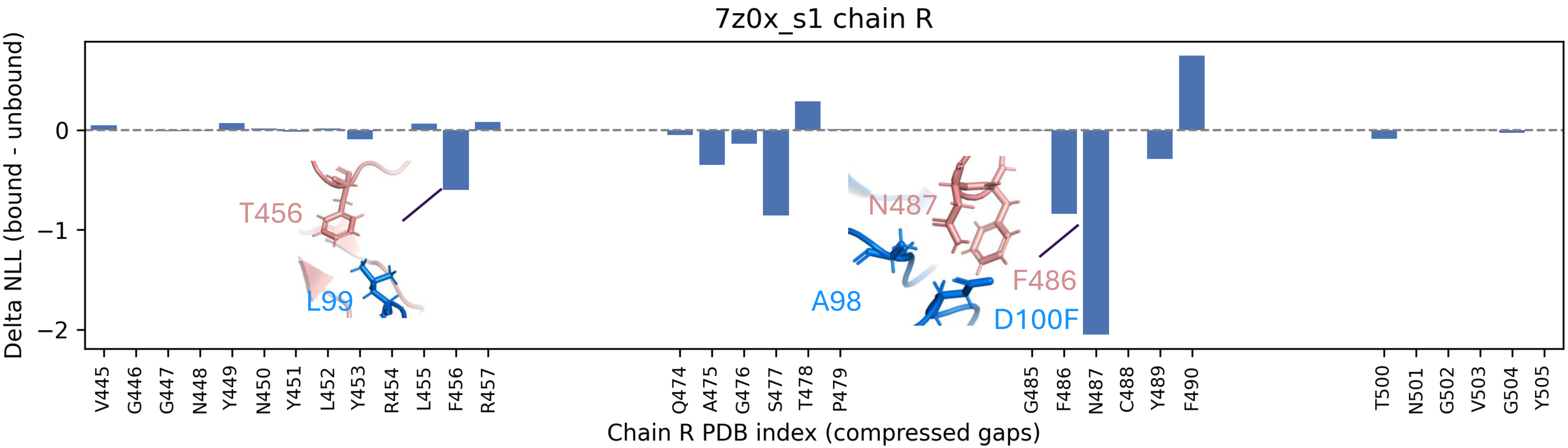
- $\Delta\text{NLL} < 0$: Key contact residue
- Residue is more probable in the bound structure than the unbound state.
- Equivalently binding context make NN “more confident” the observed amino acid. on that position is compatible given the Structure$_{AB}$

- $\Delta\text{NLL} < 0$: Potential modification spot
- *vice versa,* Unbound makes residue more favorable.

- Note this is equivalent of:
- $\Delta\text{NLL} = -\log p_{\text{bound}} + \log p_{\text{unbound}} = \log \frac{p_{\text{bound}}}{p_{\text{unbound}}}.$

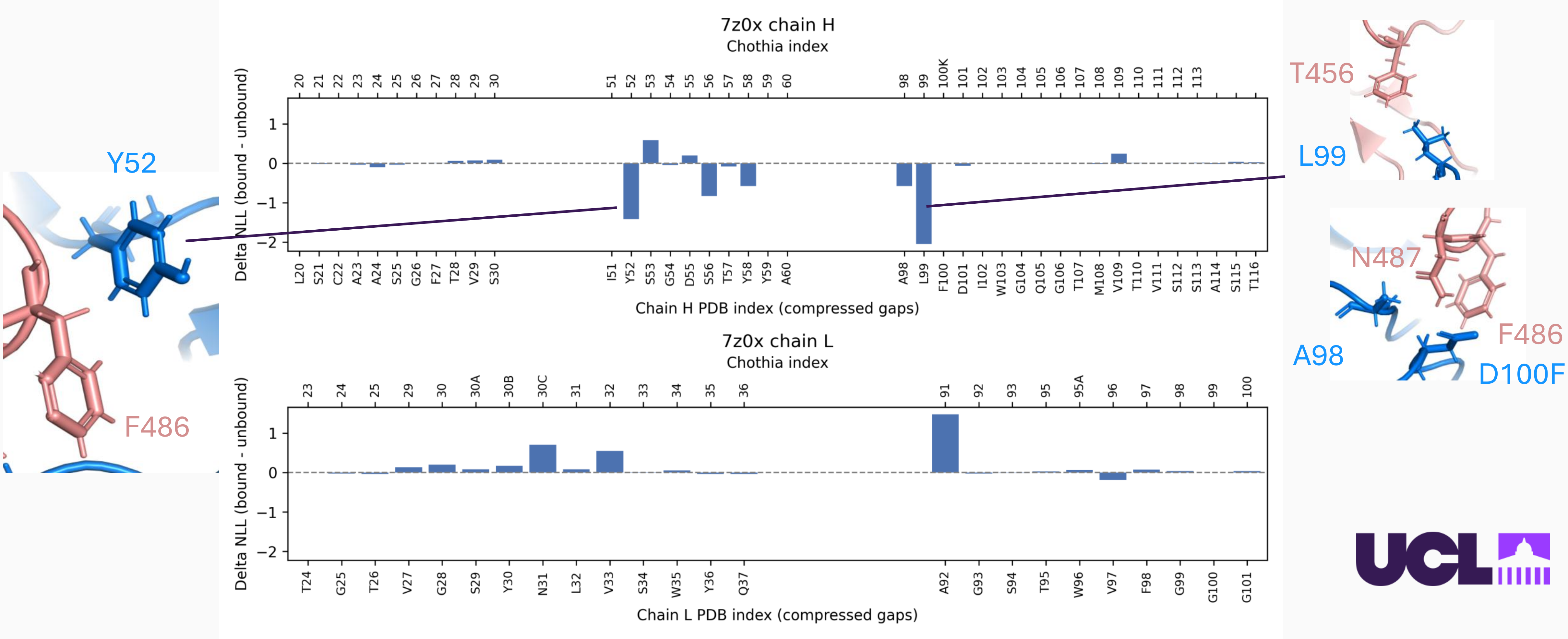
Result: S1–Ab complex (S1)

ΔNLL correlates with key contact residues
key contact residues: [478, 484, 486, 487]

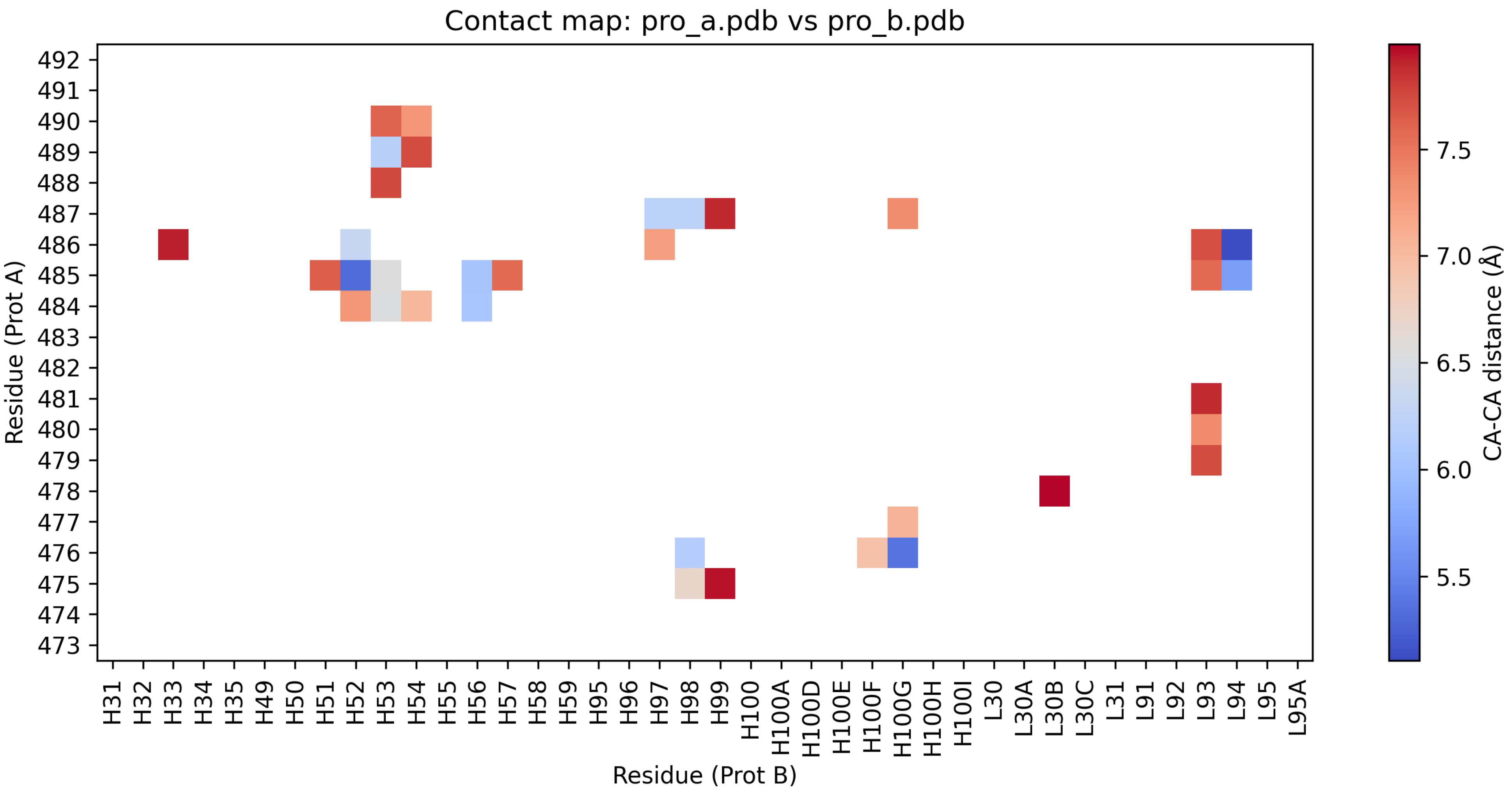


Result: S1–Ab complex (Ab)

S1-Ab complex

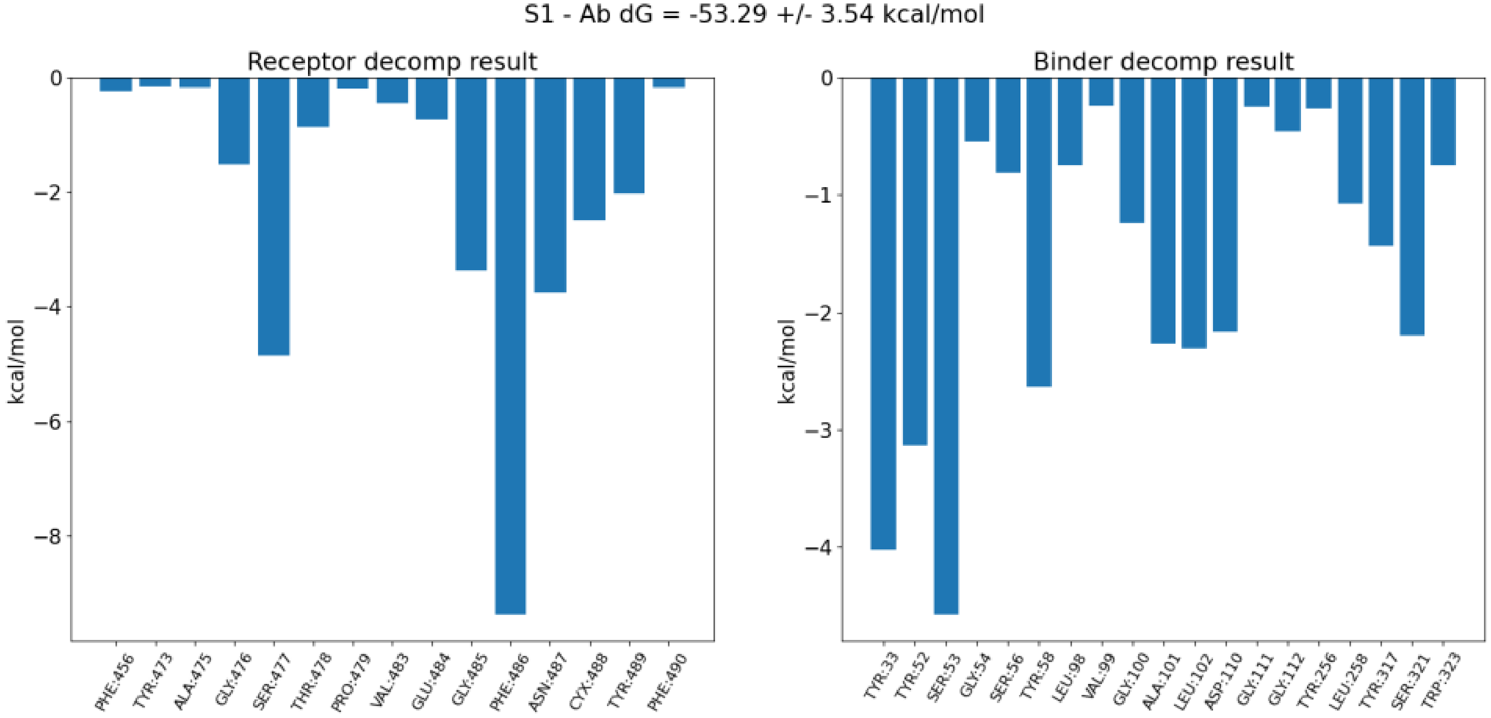


Result in line with contact map



Prev result:

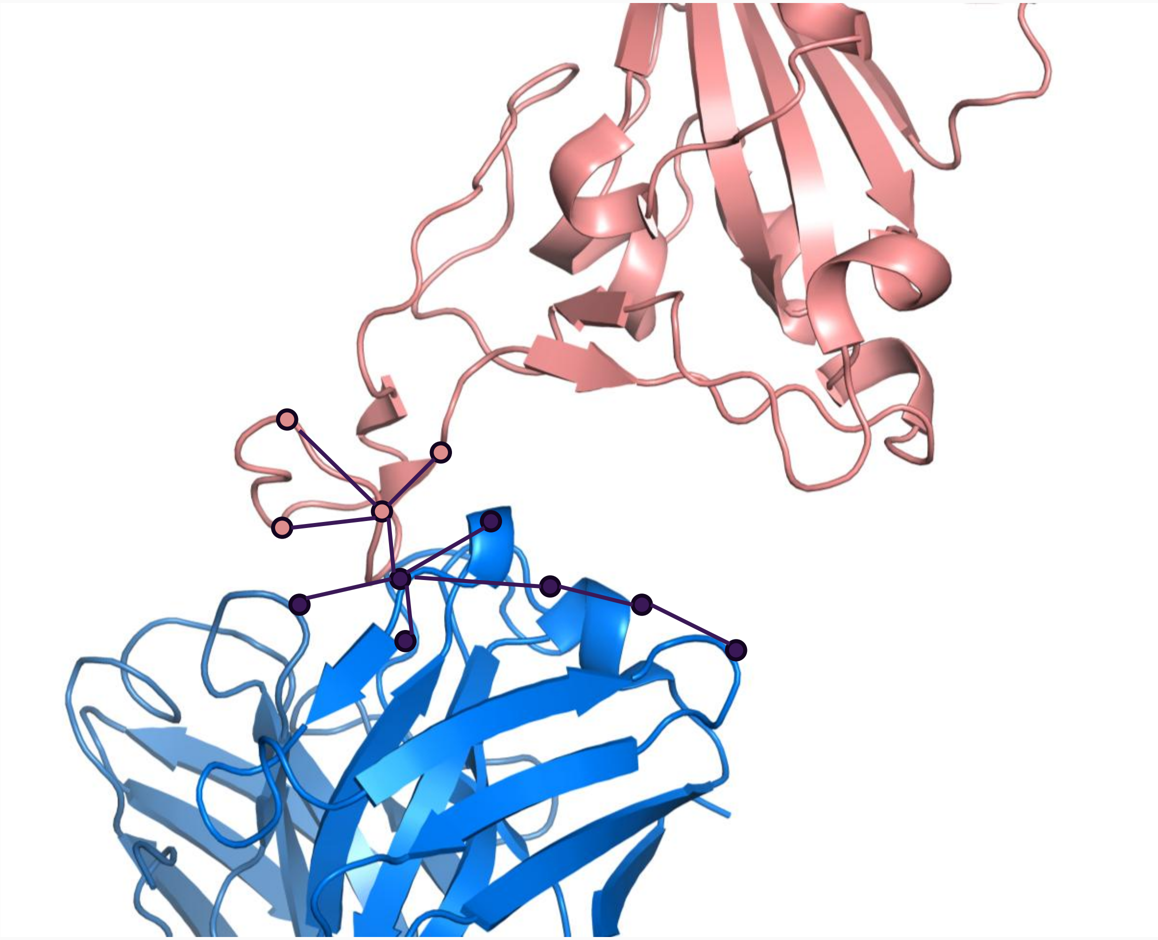
MMGBSA per-residue decomposition S1-Ab



Suggest the residue 486 on S1 protein contribute the most (~-9 kcal/mol), while the Ab has several key residues Tyr 33, Tyr52 and Ser53 etc..

Why does ΔNLL work?

- ProteinMPNN relies on graph connection around local structure.
- We are essentially using bound/unbound local structure to estimate “natural-ness” of the sequence given its structural constraint.
- The difference helps **isolate the effect from “presence of binder”** and constraint from protein itself.



Pro & Cons

- Pro
- Fast – 100 sampling on scores with less than 1 minute for target/binder sequence.
- Correlate well with Biophys/Experimental result
- Identify both key contact and potential modification spot.

- Con
- Not energy-based!
- Results essentially based on structure well-ness.
- Requires precise bound structure.

END
Thanks for listening!