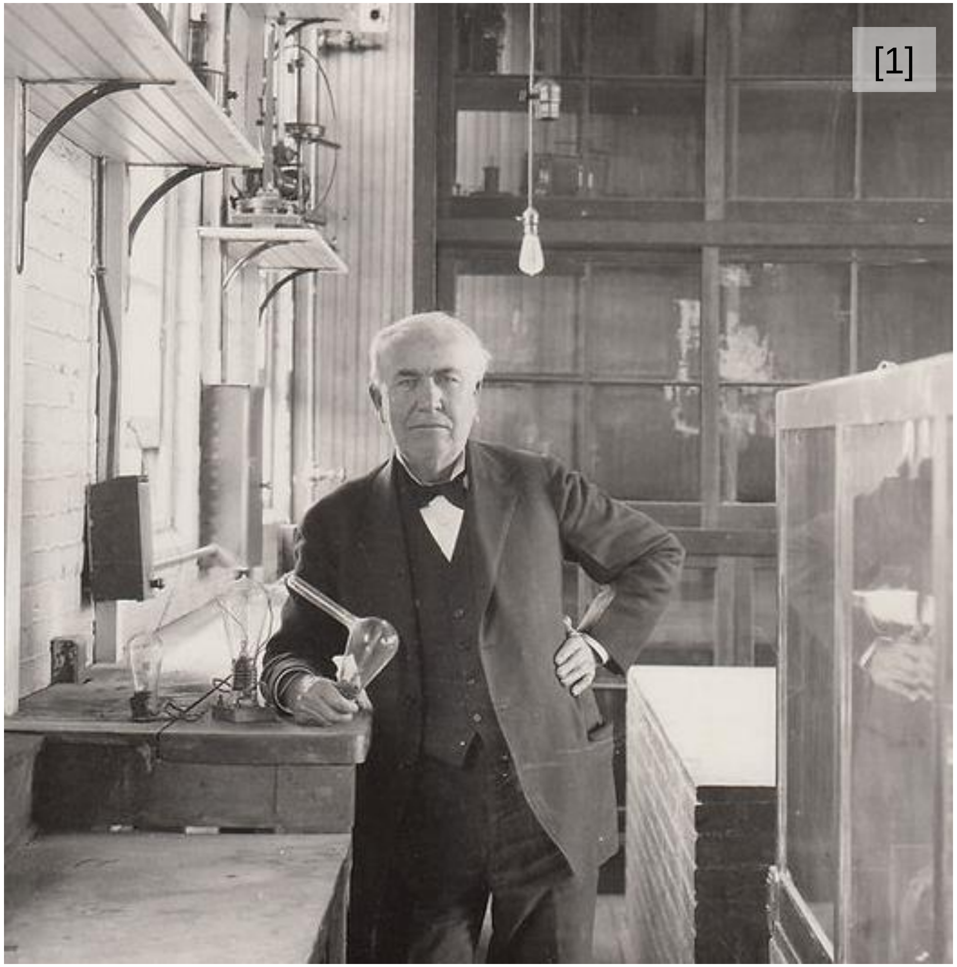
Using RF*diffusion* for Inverse Design of Proteins

Andrew Pike
CHEM101
June 3, 2024

The Edisonian Approach

“I have not failed. I've just found 10,000 ways that won't work.”
-Thomas Edison [1]

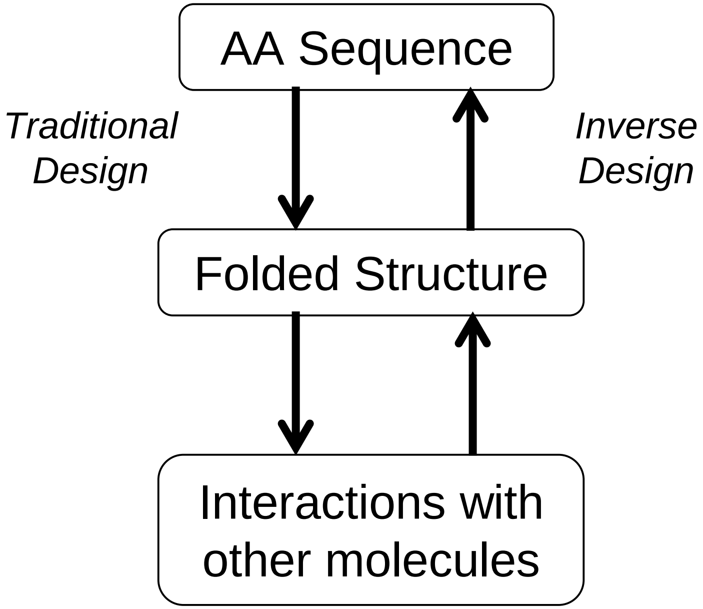
How can we design proteins more intelligently than Edison did lightbulbs?



[1] https://www.thomasedison.org/edison-quotes

RF*diffusion*

- Generate Protein structures from a specified target
- Workflow:
  - Rf*diffusion* → Generate backbone structure
  - ProteinMPNN → Determine AA sequence
  - Alphafold → Verify structure folds correctly
- Diffusion Models were previously used for image enhancement and generation



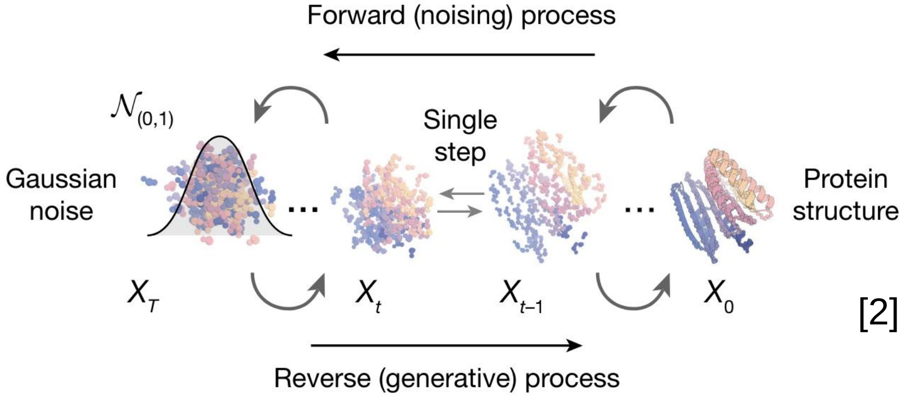
AA Sequence

Traditional Design

Inverse Design

Folded Structure

Interactions with other molecules

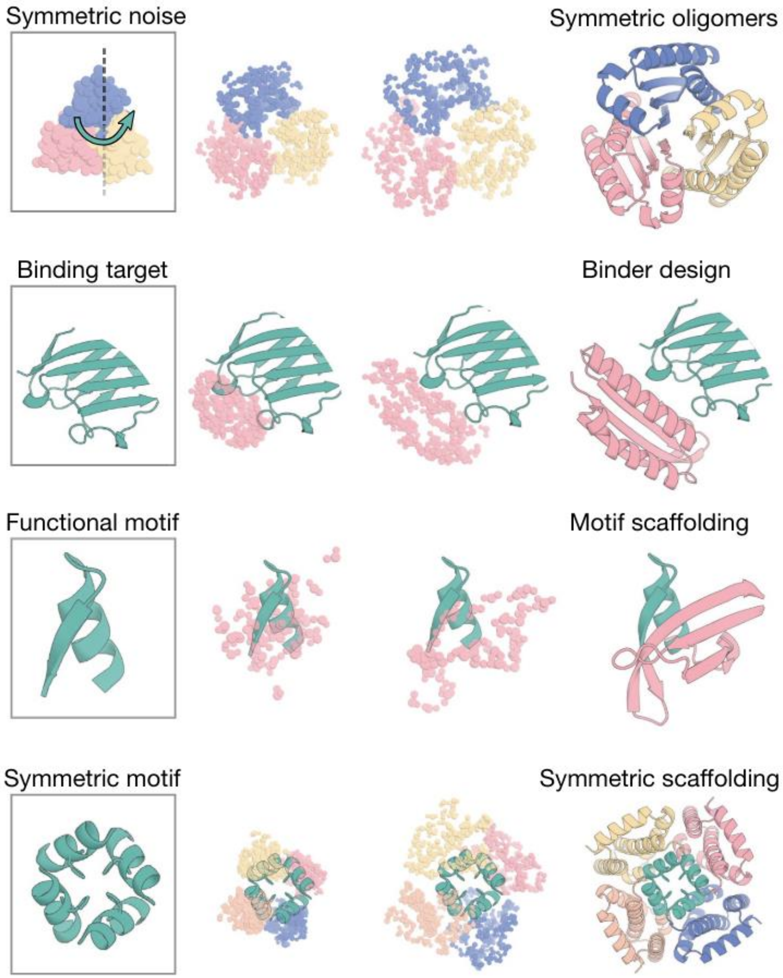


[2]

[2] Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., … Baker, D. (2023). De novo design of protein structure and function with RFdiffusion. *Nature*, *620*(7976), 1089–1100. https://doi.org/10.1038/s41586-023-06415-8

Rf*diffusion* (contd)

- Generate a set of *unrelated* structures
- If we start with random noise, how do we design anything purposefully?
- Give the model a variety of inputs
  - Sort of like providing an image that is partially blurry



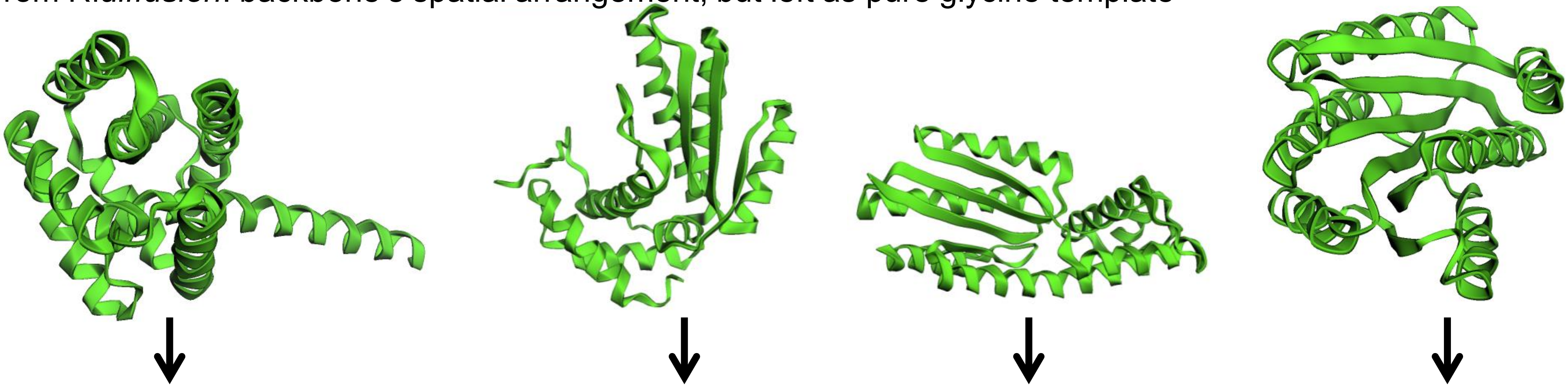
[2]

[2] Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., … Baker, D. (2023). De novo design of protein structure and function with RFdiffusion. *Nature*, *620*(7976), 1089–1100. https://doi.org/10.1038/s41586-023-06415-8

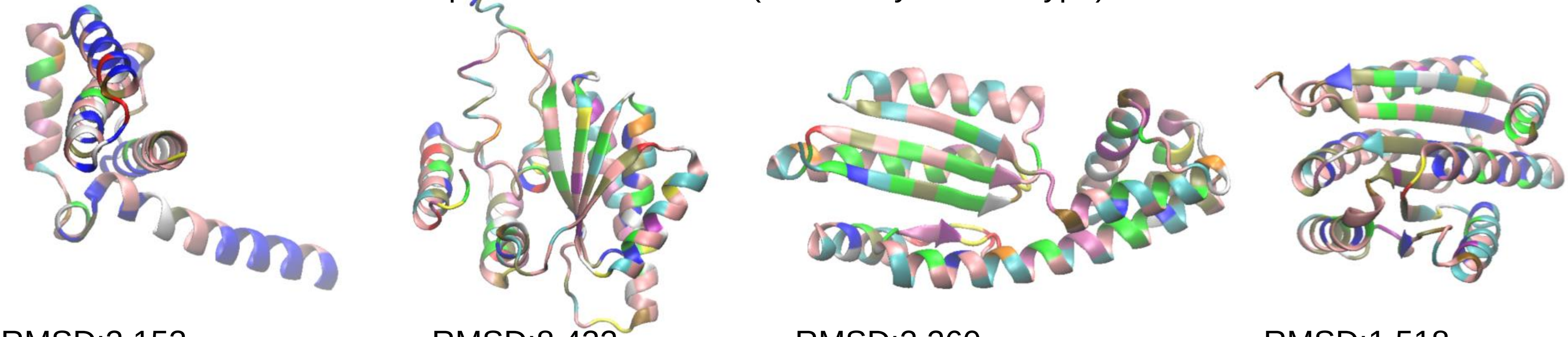
Let’s try it

- Generating some unconditional monomers 200AA in length [3]:

From Rf*diffusion*: backbone’s spatial arrangement, but left as pure glycine template



From ProteinMPNN and AF: AA Sequence and structure (colored by residue type)

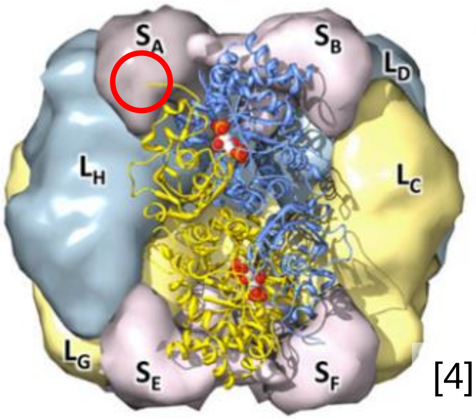


RMSD:3.153 RMSD:8.432 RMSD:2.260 RMSD:1.518

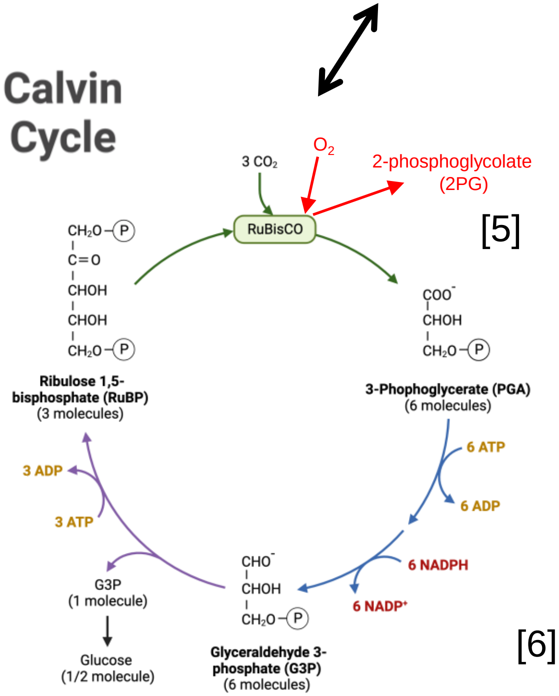
Yup, those sure do look like proteins!

[3] https://colab.research.google.com/github/sokrypton/ColabDesign/blob/v1.1.1/rf/examples/diffusion.ipynb

Rubisco



[4]



[5]

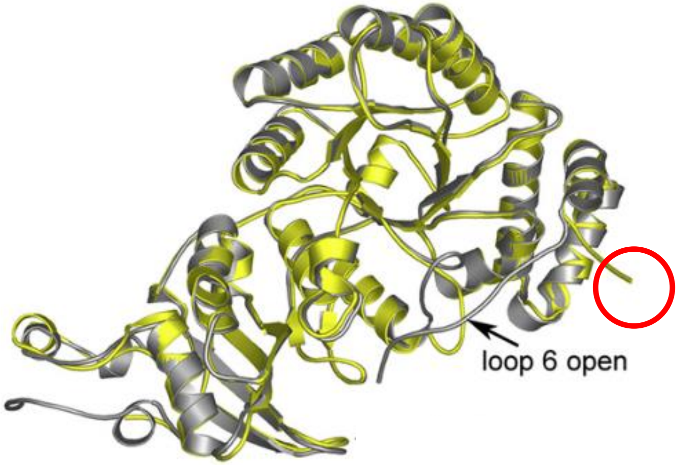
[6]

- Earth's most abundant protein and 30% of soluble protein in plants [4]
- 2 active sites are formed at the interface between 2 large subunits
- Rubisco is slow and has poor selectivity
  - ~30% energy is wasted to remove 2PG [5]
- Evolution is stuck in a local minimum
  - Most mutations are destabilizing [4]

[4] Studer, R. A., Christin, P. A., Williams, M. A., & Orengo, C. A. (2014). Stability-activity tradeoffs constrain the adaptive evolution of RubisCO. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(6), 2223–2228. https://doi.org/10.1073/pnas.1310811111
[5] Erb, T. J., & Zarzycki, J. (2018). A short history of RubisCO: the rise and fall (?) of Nature's predominant CO2 fixing enzyme. *Current Opinion in Biotechnology*, *49*, 100–107. https://doi.org/10.1016/j.copbio.2017.07.017
[6] https://slcc.pressbooks.pub/collegebiology1/chapter/the-calvin-cycle/ (modified)

Improving rubisco

- Carboxy terminus closes over the active site to allow the reaction to occur
- “…the disruption of the contacts of residue 473 (with Arg134 and His310) in the mutant enzymes causes destabilisation of the under-lying loop 6.” [7]
- Design a binder that connects to loop 6 (134), the carboxy terminal (473), then back to loop 6 (310)



*rbc*L from spinach (gray) and *rbc*L from *Halothiobacillus neapolitanus* (sulfur oxidizing chemoautotroph) [7]

Input to Rf*diffusion*:

```
./RFdiffusion/run_inference.py
inference.output_prefix=outputs/test
inference.num_designs=4
inference.input_pdb=outputs/test/rubisco.pdb
diffuser.T=50
'contigmap.contigs=[5-15/A133-135/5-15/A472-
474/5-15/A309-311/5-15]’
 inference.dump_pdb=True
inference.dump_pdb_path='/dev/shm'
```

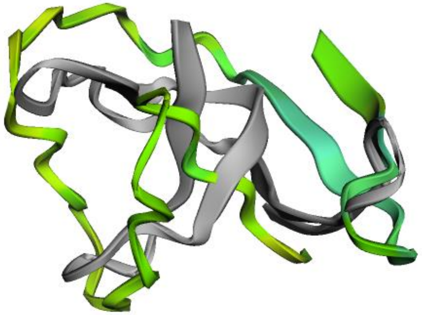
[7] Andersson, I., & Backlund, A. (2008). Structure and function of Rubisco. *Plant Physiology and Biochemistry*, *46*(3), 275–291. https://doi.org/10.1016/j.plaphy.2008.01.001

[8] https://www.rcsb.org/structure/5iu0

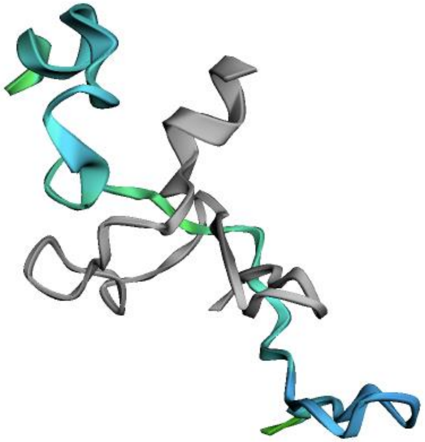
Improving Rubisco

- Same workflow as before, but with Rubisco to guide generation
- AF matches to RF*diffusion* more poorly than unconditional generation

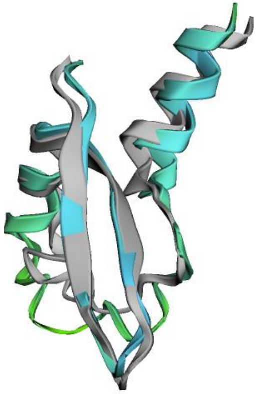
RMSD: 10.8



RMSD: 9.9



RMSD: 11.2

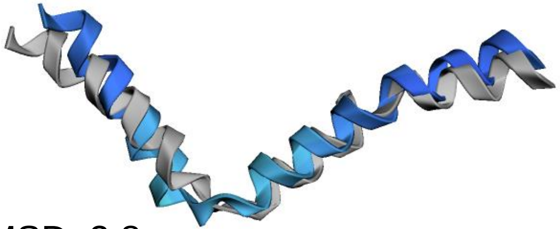


RMSD: 6.0

```
'contigmap.contigs=[5-15/A133-135/5-
15/A472-474/5-15/A309-311/5-15]’
```

↓

```
'contigmap.contigs=[5-15/A134-134/5-
15/A473-473/5-15/A310-310/5-15]’
```

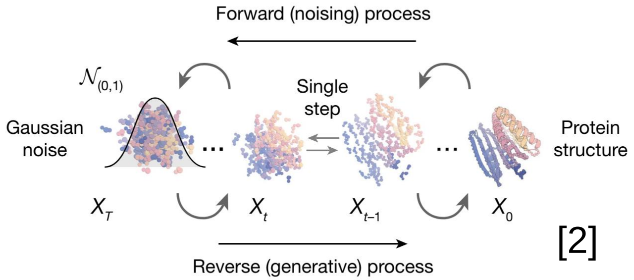


RMSD: 2.8

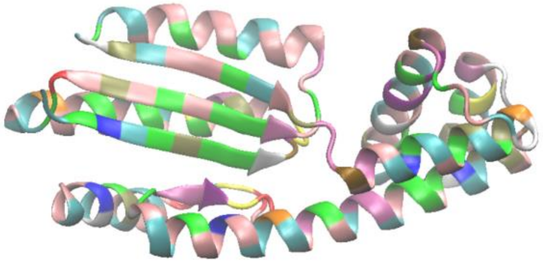
No direct indication of catalytic activity, but these could be interesting prototypes for further calculations

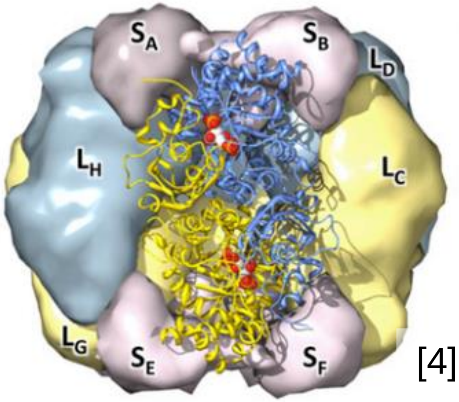
Conclusions

- Rf*diffusion* is a step towards inverse design of proteins
- Coupled with ProteinMPNN, structures accurately match those predicted by AlphaFold
- Generated several binders to interact near the active site of Rubisco
- Further computation and experiments required to fully determine their effects!
    - Nevertheless an example of what Rf*diffusion* is capable of
    - An interesting way to study Rubisco *in silico*
- Expert guidance can be incorporated to refine Rf*diffusion*



[2]





[4]