Application of diffusion models to rewind the prebiotic accretionary formation of ribosomes (RiboDiffusion)

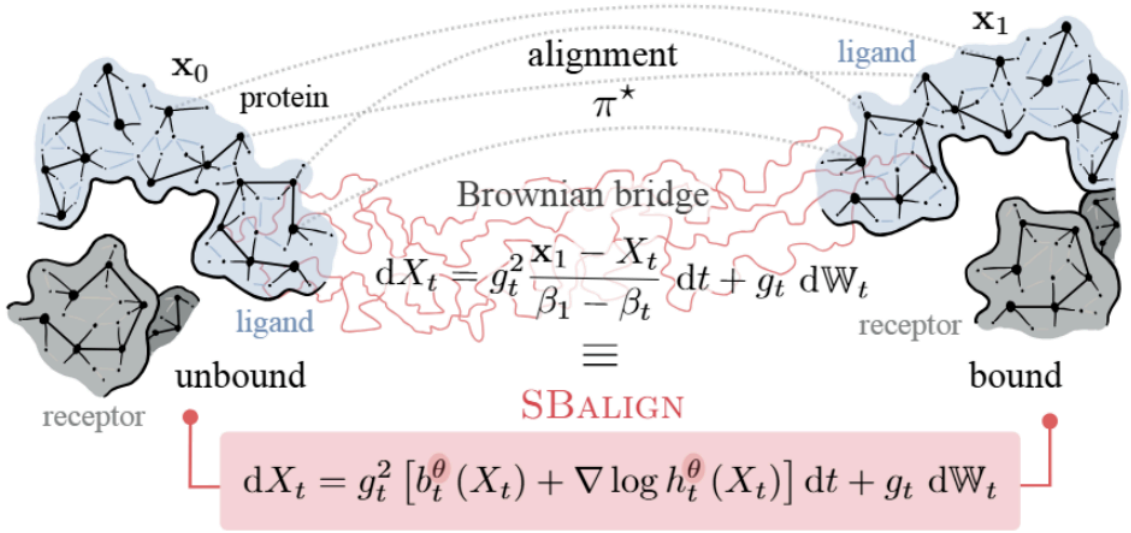

Figure 1 (reference2)

A ribosome is a fundamental, ancient molecular machine (> 300.000 atoms) essential to all known life. It is primarily composed of ribosomal RNA and ribosomal proteins, forming a partnership that dominates biology. Functionally, the ribosome is responsible for translation, the process of converting genetic information encoded in messenger RNA (mRNA) into coded protein sequences. This process is universal across all living systems.

The ribosome consists of two main components: the small ribosomal subunit (SSU), which is responsible for decoding the mRNA sequence, and the large ribosomal subunit (LSU), which catalyzes peptidyl transfer, the chemical reaction that links amino acids together to form a protein chain. It is one of the oldest existing biological systems, with its essential structural foundation, the universal common core, finalized by the time of the Last Universal Common Ancestor (LUCA) around 3.8−4.2 billion years ago. The structure of the ribosome itself, particularly the layered organization of its rRNA and protein components, serves as a molecular fossil, preserving information about its long evolutionary history of incremental accretion. Despite its complexity and essential role, the ribosome is structurally robust and functionally resilient, making it very hard to render completely nonfunctional.

The period before LUCA is referred to as the root phase of the tree of life. During this time, the ribosome underwent significant development, characterized by a process of discontinuous accretion. This means that ribosomal structure and function were built up incrementally and iteratively without substantial remodeling of pre-existing basal structures. The ribosome has recorded its history by accreting rRNA, ribosomal proteins (rProteins), and inorganic cations. According to the accretion model, six phases of ribosomal evolution took place during the root of the Tree of Life. Each phase involved the addition of Ancestral

Expansion Segments (AESs in the LSU and aes's in the SSU) and associated protein segments.

After LUCA, the ribosome entered stasis and remained in that state for several billion years. Bacterial ribosomes never left stasis. Archaeal ribosomes have remained near stasis, with the notable exception of the superphylum Asgard, which has accreted rRNA post-LUCA. Archaeal rRNAs are slightly larger than bacterial rRNAs on average, primarily due to archaeal-specific expansion segments (μ-ESs) inserted onto the surface of the common core. Eukaryotic ribosomes in some lineages emerged from stasis around 2 billion years ago and entered a dynamic phase of growth, which is ongoing and accelerating in some metazoan lineages. Eukaryotic ribosomes are much larger than the common core, with more complex structures. This size increase is primarily due to the iterative insertion of small RNA fragments, known as Expansion Segments (ESs in the LSU and es's in the SSU), into universally conserved sites on the subunit surfaces.

In summary, the ribosome's history shows a rapid, complex evolution before LUCA to finalize the universal common core, followed by a long period of stasis in bacteria, slight accretion in some archaea, and significant, accelerating accretion in eukaryotes.

The accretion of ribosomes bears some resemblance to diffusion models in AI (figure 2). Given the success of diffusion models in exploring the structural space of biomolecular complexes (reference 5), and recent suggestions for their use in studying formation pathways, (reference 6) could these models be applied to investigate the possible prebiotic accretionary paths of ribosomes? **This raises the question of whether the proposed ribosome history represents the singular path or one of many possibilities, and how diffusion models might help elucidate this.**

Reference 2 introduces Diffusion Schrödinger Bridges (DSBs) as a framework for recovering stochastic dynamics via marginal observations at different time points. Their novel approach, SBALIGN, is specifically designed to utilize the structure of aligned data, which naturally arises in many biological phenomena.

Examples of biological applications include predicting conformational changes in proteins and the temporal evolution of cellular differentiation processes (which are a kind of accretion model). In protein docking, this involves modeling the stochastic trajectory from an unbound to a bound structure (figure 1). For cell differentiation, it's about reconstructing the dynamics between observed states at different times.

The accretionary evolution of the ribosome before LUCA, as discussed in our previous conversation, describes a process where the ribosome built up its structure incrementally over time, from smaller protoribosomes to the complex universal common core. This can be viewed as a kind of stochastic trajectory through a space of possible molecular structures and functions.

If we had representations of hypothesized ancestral ribosomal states at different points in this accretionary history (analogous to the x0 and x1 protein structures or cell states), a DSB framework could potentially be used to model the transitions and dynamics that connect these states.

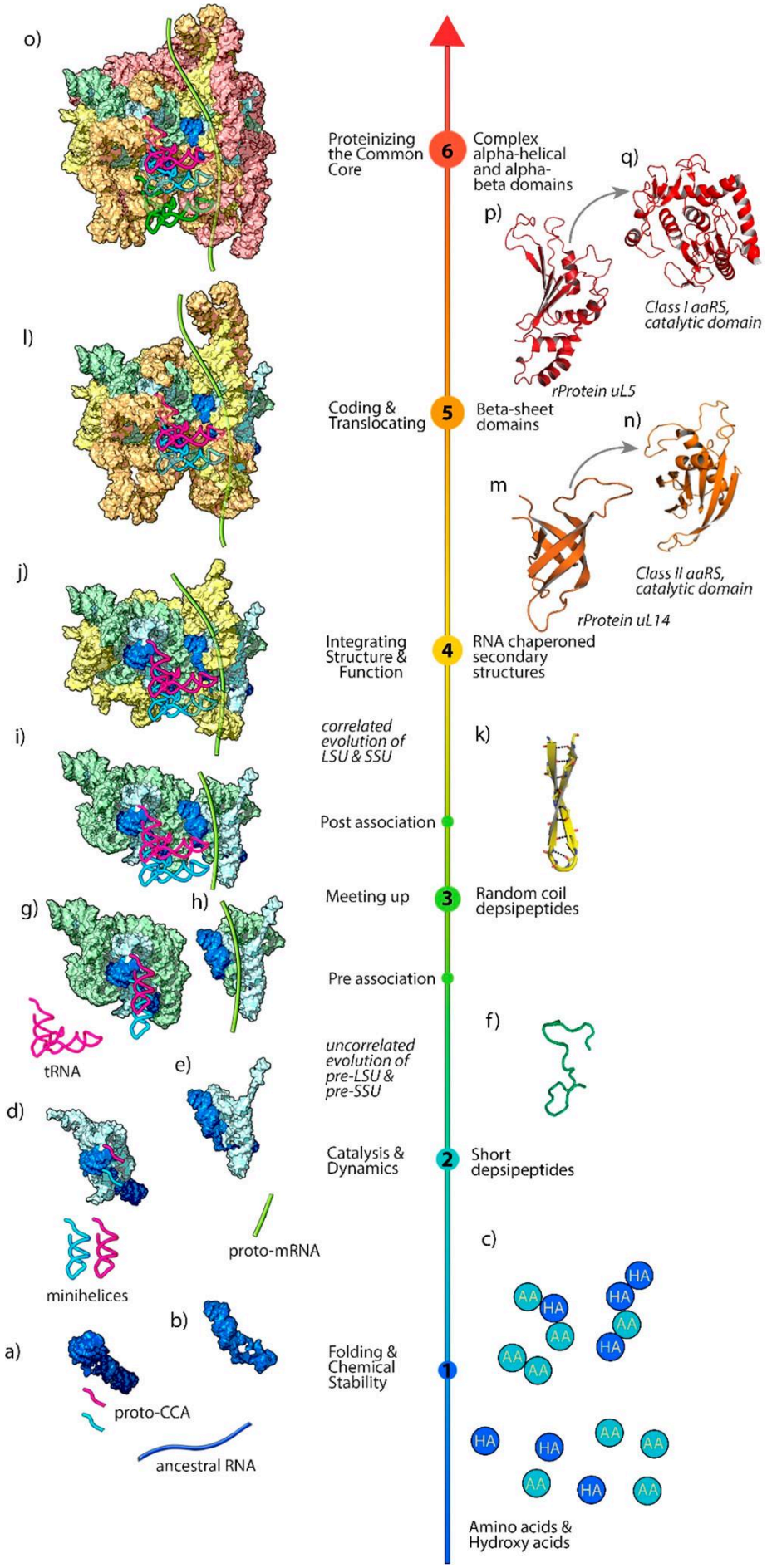

Figure 2 (reference 1)

RiboDiffusion 3

"Rewinding" the evolution could conceptually correspond to leveraging the backward process of the DSB model, sampling trajectories that move from a later, more complex state (like the LUCA common core) back towards earlier, simpler protoribosomal states, similar to how SBALIGN reconstructs the process from x1 back to x0 or along the intermediate trajectory. The challenge lies in defining and obtaining the "aligned data" representing these deep evolutionary states.

Reference 3 establishes a fundamental connection between diffusion models and evolutionary algorithms, noting that both generate high-quality samples via iterative refinement on random initial distributions. They propose using deep learning-based diffusion models (DM) as generative models within evolutionary tasks.

The pre-LUCA ribosome evolution was an evolutionary process characterized by the iterative accretion of rRNA and proteins, leading to increased functional complexity. This process resulted in the "birth" of increasingly sophisticated protoribosomes. A diffusion model could potentially be trained to act as a generative model for ancestral protoribosomal structures or sequences, perhaps viewed as "gene expressions" that encode aspects of protoribosome function and structure.

"Rewinding" the evolution in this context might involve using such a trained DM to sample hypothetical earlier forms of protoribosomes based on properties of the LUCA core to bias the generation towards properties expected of ancestral forms, like smaller size or specific structural motifs like the PTC or DCC cores. This would be less about reconstructing a single historical path and more about exploring the space of plausible ancestral forms under the generative principles learned by the model. The iterative denoising process within the DM itself shows a resemblance to step-wise biological processes.

References

1. **Root of the Tree: The Significance, Evolution, and Origins of the Ribosome**
   Jessica C. Bowman *et al*. *Chemical Reviews* **2020** 120 (11), 4848-4878
   DOI: 10.1021/acs.chemrev.9b00742
2. **Aligned Diffusion Schrödinger Bridges**
   Somnath, V. *et al arXiv* **2023**.
   https://doi.org/10.48550/arxiv.2302.11419.
3. **Learning Diffusion Bridges on Constrained Domains**
   Liu, X. *et al*. *ICLR **2023***.
   https://openreview.net/forum?id=WH1yCa0TbB
4. **Heuristically Adaptive Diffusion-Model Evolutionary Strategy**
   Hartl, B. *et al* M. *arXiv* **2024**.
   https://doi.org/10.48550/arxiv.2411.13420.
5. **Generalized biomolecular modeling and design with RoseTTAFold All-Atom**.
   R. Krishna et al. *Science* **2024**, eadl2528.
   https://www.science.org/doi/10.1126/science.adl2528
6. **Protein structure generation via folding diffusion.**
   K. E. Wu et al. *Nat. Commun.* **2024,** 15, 1059.
   https://www.nature.com/articles/s41467-024-45051-2