ICML Oral

Towards Sub-Second Molecular Docking as a Structural Primitive

A Quantized Consistency Diffusion Framework for Agent-Centered Drug Discovery

Latency determines whether structure can stay inside the reasoning loop.

Low-latency primitives make the scientific AI workbenches executable.

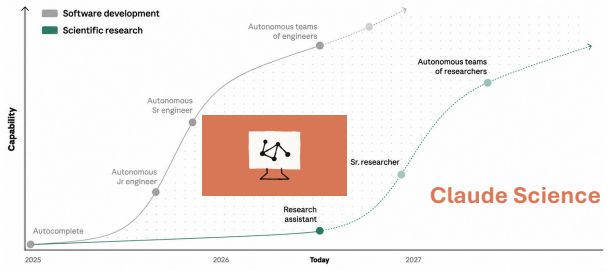
上海科技大学 ShanghaiTech University

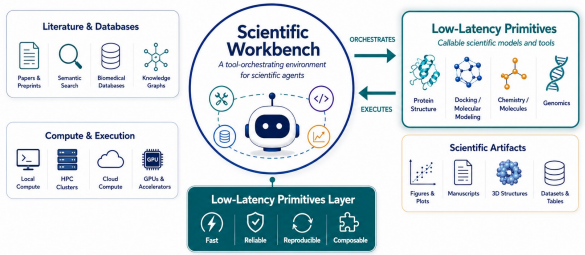
Scan to connect on WeChat

Email: zhangkexin@proteindance.cn

1 Why now? Agentic discovery

The climb in life sciences runs through the real world





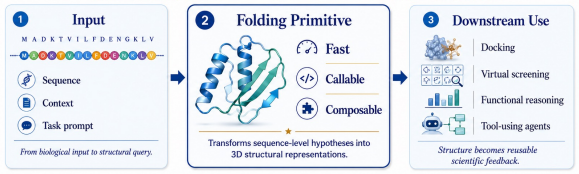
Scientific workbenches orchestrate. Low-latency primitives execute.

Low-latency primitives make the scientific workbench executable.

Folding as a Structure Primitive

Structure prediction as a callable capability for scientific agents

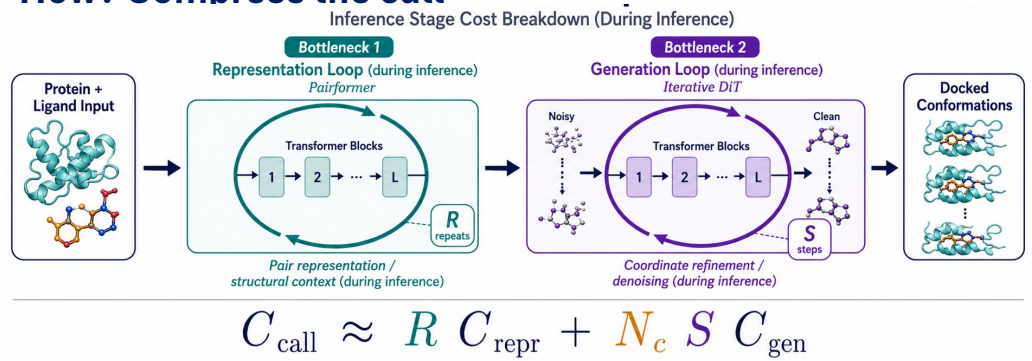
A fast structure primitive turns sequence into actionable structural feedback.



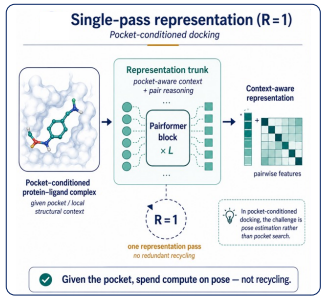
Question → Fold → Analyze → Revise

Folding is not only a prediction task — it is a low-latency structural primitive for the reasoning loop.

2 How? Compress the call



$$C_{\text{call}} \approx R\, C_{\text{repr}} + N_c\, S\, C_{\text{gen}}$$

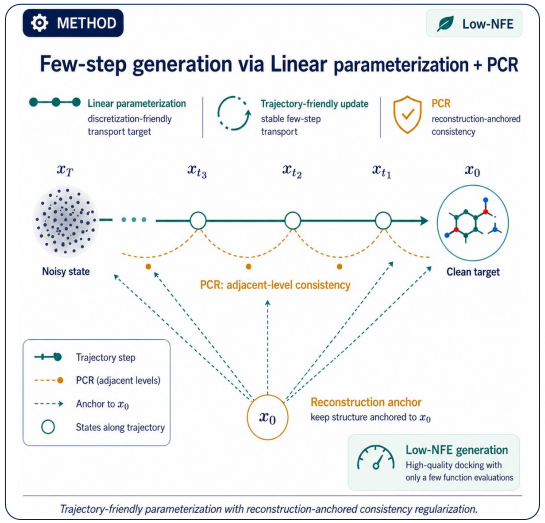


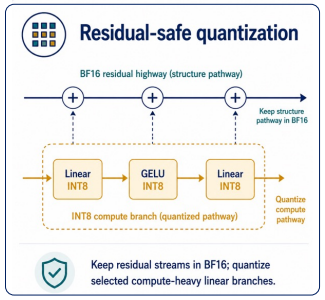
Unified √NSR

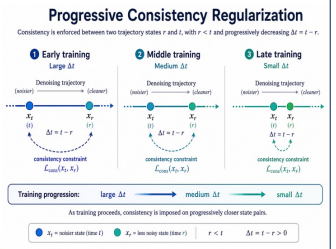
$$z_\lambda = \frac{x_t}{\sqrt{\alpha^2(t)+\sigma^2(t)}} = \frac{x_0+\lambda\epsilon}{\sqrt{1+\lambda^2}}$$

Schedule-independent normalized noisy input.

***NSR** is the common coordinate that aligns noisy inputs, timesteps, discretization, and loss weighting across any diffusion schedules.*







AI workbenches orchestrate; low-latency primitives execute.

3 So what? Next frontier

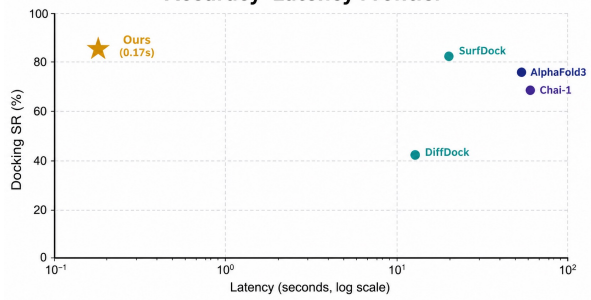
0.17s
Ultra-fast Inference
- 5 conformations
- Single H20 GPU

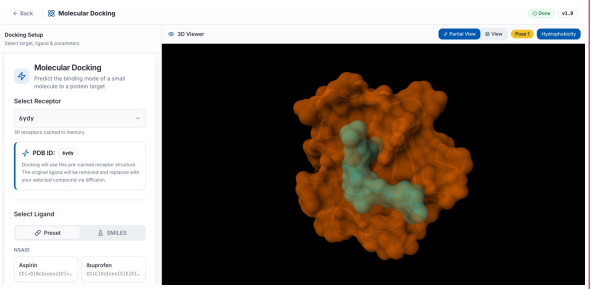
>300×
Speedup over AlphaFold3
- Dramatically faster
- Massively more efficient

84.1%
Pocket-Conditioned Docking on PoseBusters
- PoseBusters benchmark
- High-accuracy pocket-conditioned docking

49.1%
Blind Docking on PoseBusters
- PoseBusters benchmark
- Robust blind docking performance

Accuracy–Latency Frontier





From product to primitive.

*Our model is not only an online docking SaaS or model-as-a-service API; it is a low-latency **Bio Token** for agentic science. It turns molecular docking from a prediction endpoint into a reusable structural primitive — callable by AI workbenches, composable by scientific agents, and executable inside real-time discovery loops.*