nature biotechnology

Review article

https://doi.org/10.1038/s41587-024-02133-2

Sparks of function by de novo protein design

	Class	Task	Method/Process	Examples
Classical approaches	*Function → structure*	Functional motif scaffolding	Functional motif → protein containing this motif	TopoBuilder (*54*), RFdiffusion (motif scaffolding) (*20*), RFjoint (*55*)
		Functional motif generation and docking	Binding target → field of possible interactions	RIFgen (*60*), COMBS (*123*), Sculptor* (*23*), MaSIF (*64, 65*)
		Zero-shot interaction generation[a]	Binding target → binding protein	RFdiffusion (protein targets) (*20*), RFAADiffusion (small molecule targets) (*21*)
	Structure	Structure generation (fragment)	Syntax → Structure	RosettaRemodel (*69)*
		Structure generation (repurposed structure prediction networks)	Sequence → optimization → full-atom structure	Hallucination (*56, 101*), RFjoint (*55*), Frank (*80*), Verkuil (*119*), Hie (*122*)
		Structure generation (non-diffusion)	Noise → backbone	Anand (*87*), SCUBA (*74*), Ig-VAE* (*23*)
		Structure generation (diffusion / flow)[b,c]	Noise → backbone	RFdiffusion (*20*), RFAADiffusion (*21*), ProteinSGM (*93*), Chroma (*94*), Protpardelle (backbone) (*95*), Genie (*24*), FoldingDiff (*25*), Framediff (*26*), FoldFlow (*27*), FrameFlow (*28*), LatentDiff (*29*), PVQD (*30*)
	Structure → sequence	Local structure-guided	Local structure → local sequence	RosettaDesign (physics-based potential) (*15*), ALP (*102*), ProteinMPNN (*103*), ProteinSolver (*31*), Structured Transformer (*32*), PiFold (*33*), Grade-IF (*34*)
		Global structure encoding (encoder-decoder)	Full structure → full sequence	ESM-IF (*35*), ABACUS2 (*36*), ABACUS-R (*37*), ProstT5 (*38*), SaProt (*39*), MIF-ST (*104*)
Other approaches	*Sequence*	Sequence generation	Noise/prior → full sequence	NOS/Lambo (*40*), ProteinGAN*(*41*), Greener (*42*), Progen2 (*112*), ProtGPT2 (*113*), EvoDiff (*114*)
	Sequence & structure	Co-design structure and sequence[d]	Noise → sequence & structure / MCMC / iterative refinement	Hallucination (*56, 80, 101*), RFjoint (*25*), Verkuil (*119*), Hie (*122*), ProteinGenerator (*127*), Protpardelle (all-atom) (*95*), Jin* (*43*), AbDiffuser* (*44*), Luo* (*45*)

Table 1: Approaches to de novo protein design.
Underlined methods include experimental validation. *Fold/family-specific methods.

[a] "Zero-shot" typically refers to model generalization to new tasks which have not been seen during training (*46*). Here we use it to refer to prediction/generation of new binders when no successful binders are used to guide solutions, though in practice the model has been trained on binder-target pairs and it is not uncommon that binding targets have been seen previously during training.

[b] Current protein diffusion models also condition on the protein length (in addition to noise). Prior distributions on length are typically uniform or task-specific.

[c] Diffusion models possess a natural relationship to flows as both are often implemented as neural ODEs (*47*). Indeed, diffusion under the probability flow ODE is a form of continuous normalizing flow, allowing exact likelihood computation and latent variable inference (*118*), and further work to translate the efficient training and performance of diffusion to flows has been explored through flow matching approaches (*48*, *49*). Both diffusion and flow models can be viewed as special cases of stochastically interpolating models, a general framework for mappings between arbitrary pairs of distributions which offers additional flexibility over standard diffusion, such as not requiring a Gaussian prior (*50*). Flow-based protein generation was first suggested by Chroma and implemented by Protpardelle (*94*, *95*), with further development in FoldFlow and FrameFlow (*27, 28*), but extension of non-diffusion stochastic interpolants to design applications remains nascent. Some of these explorations relate to using non-Gaussian priors which have a coupling with the target distribution (i.e. paired data), such as sampling conformations from structure or binding complexes from monomers (*27*, *51*).

[d] Many "joint" methods and compositions of methods are presented as models of the joint distribution of structure and sequence, and can indeed be posed as such: they might be able to ascribe a probability, energy, or density to a set of structure and sequence variables; or more stringently, they might admit sampling of (structure, sequence) pairs which are mutually consistent, whether simultaneously, sequentially/ancestrally, in a Gibbs-based fashion, or otherwise. We suggest that to be maximally effective, a joint co-design model should possess these capabilities in addition to methods for marginalizing and conditioning the joint distribution (i.e. conduct structure and sequence generation independently, as well as structure prediction and sequence design).

Methods	PDB ID
Library screening	8H7C
Rational design	7BEY, 6Z0L, 6Z0M, 6REN, 6ZT1
Rosetta + Library screening	7BWW, 6OHH
Rosetta fragment assembly from blueprints	8BL6, 7SKP, 7SKO, 7SKN
Crick equations + Rosetta HBNet + RosettaDesign	6MSQ, 6MSR, 8GL3, 6N9H, 6NAF
Rule-Based + Database Fragment Search	6MCT, 6MQU, 6MPW, 6MQ2, 8DPY
Rule-Based + Negative Design in Rosetta	6X9Z
Database Interactions Search	6W70, 5HRZ
TopoBuilder + Rosetta FunFolDes + Library screening	6YWD, 6YWC
RoseTTAFold Joint Inpainting	8DT0
AF2 Hallucination + ProteinMPNN	8FJG, 8FJF, 8FJE, 8CYK, 8OYY
trRosetta Hallucination	7K3H, 7M0Q
Rosetta Remodel	8GAA
Rosetta Remodel + RosettaDesign	6NX2, 6NXM, 6NY8, 6NYE, 6NYI, 6NZ3, 6NZ1, 6NYK
Rosetta Remodel + ProteinMPNN	8EOX, 8EOZ
Custom RosettaScripts	8FBI, 8FBN, 8FBJ, 8FBK, 8E55, 8E1E (DegreaserMover), 7JH5, 7CBC (GraftSwitchMover)
Kinematic Loop Closure	6UD9, 6UFU, 6UF7, 6UDW, 6UF8, 6UFA, 6UDR
Crick equations + Kinematic Loop Closure	8EK4
Rule-Based + Rotamer Interaction Field	6D0T, 6CZI, 6CZH
Rule-Based	8BFD, 8A09
RPXDock	8FWD
RosettaDesign	6VFK, 6VFH, 6VFI, 6VFJ, 6VL6, 6VEH
WORMS	6XNS, 6XT4, 6XH5, 6XSS
Rosetta SEWING	7TJL

Table 2: Methods used to design select protein structures in Fig. 6.