

Advances in Generative and Discriminative Models for *De Novo* DNA Binder Design

Paul Taewan Kim

A dissertation

submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

David Baker, Chair

Frank DiMaio

Hugh Hillhouse

Barry Stoddard

Program Authorized to Offer Degree:

Molecular Engineering

©Copyright 2026

Paul Taewan Kim

University of Washington

Abstract

Advances in Generative and Discriminative Models for *De Novo* DNA Binder Design

Paul Taewan Kim

Chair of the Supervisory Committee:

David Baker

Department of Biochemistry

In organisms across domains of life, the process of gene regulation, or the expression of information stored in the genetic code, is dominated by interactions between proteins and DNA. In particular, various families of DNA binding proteins are responsible for: scaffolding, winding, unwinding of the DNA strand; promotion and suppression of specific genes; damage detection and repair; and more. Although numerous engineering tools have been developed in recent years to manipulate DNA, these tools are derived from naturally-occurring biological matter, and are therefore limited in scope and function — at best, we can slightly modify or mix and match functions that nature has evolved. Recent advances in the field of de novo protein design, where proteins are rationally designed by humans from scratch, promises to unlock new, powerful genetic engineering tools, with novel function wholly outside of the scope of existing biological matter. In this thesis I describe efforts to advance the field of de novo protein design for DNA binding, starting with 1. a set of increasingly powerful generative machine learning

models for protein design, applied to DNA binders; 2. a computational pipeline for designing DNA binding proteins with the ability to discover sequence selective DNA binding proteins from as few as 96 designs tested in the wet lab, and 3. discriminative models for DNA specificity prediction, to enhance the ability to filter candidate DNA binding proteins.

Contents

List of Figures	6
List of Tables	8
Acknowledgments	9
1 Introduction	11
1.1 Biochemistry of Proteins and DNA	12
1.2 Protein-DNA Interactions in the Cell: Structure and Function	14
1.3 Existing Discovery Approaches to Protein-DNA Interactions	16
1.4 The Promise of De Novo Design	18
1.5 Prior Work: Glasscock et al.	19
1.6 Research Directions and Goals	20
2 Generative Models for DNA Binder Design	22
2.1 Diffusive Generative Models	22
2.2 Modeling of Proteins and DNA	26
2.3 Diffusion Models for Protein Design	30
3 De Novo Generation of Sequence-Selective DNA Binders	44
3.1 The Generative Model: RFdiffusion3	44
3.2 A Pipeline for Sequence-Selective DNA Binder Design	50
3.3 In-Silico Results	54
3.4 Experimental Characterization	56
3.5 Future Work: Latent Diffusion and Inference-Time Search	65
3.6 Discussion	68

4	Discriminative Models for DNA Binder Prediction	71
4.1	Motivation	71
4.2	Data Curation	75
4.3	Modeling	79
4.4	<i>In Silico</i> Performance	81
4.5	Further Work	84
	References	85
	Figures	107
	Tables	145

List of Figures

1	Chemical patterns that distinguish base pairs at the groove edges . . .	107
2	Four recurring DNA-binding scaffolds	108
3	Established routes to sequence-specific DNA-binding function	109
4	The RIFdock design pipeline of Glasscock et al	110
5	Designed DNA-binding proteins of Glasscock et al. are sequence-specific	110
6	The forward and reverse processes of a denoising diffusion model . . .	111
7	Structural representations used by structure prediction and design models	111
8	The alpha-carbon diffusion design pipeline and its RoseTTAFoldNA pass rates	112
9	Yeast-display sorting of the alpha-carbon diffusion library against one DNA target	112
10	Alpha-carbon diffusion designs against the HBVA target bind with high affinity and selectivity	113
11	RoseTTAFoldNA metrics of alpha-carbon diffusion designs, grouped by experimental outcome	113
12	The RFdiffusion All-Atom pipeline for DNA binder design, and its <i>in silico</i> pass rates	114
13	Input to the tip-atom RFdiffusion All-Atom model	115
14	Structure-prediction metrics as <i>in silico</i> filters, evaluated retrospectively	115
15	Outcome of the RFdiffusion All-Atom design campaign	116
16	All-atom protein design with RFdiffusion3	117
17	Global and atom-level conditioning with RFdiffusion3	119
18	Design approach	121
19	Extended depiction of the de novo DNA binder design pipeline	122
20	Schematic description of Δ_{minPAE} calculation	124
21	Experimental characterization of selectivity of designed DNA binding proteins	125
22	Correlation between experimental binding and Δ_{minPAE}	127

23	Biophysical comparison of designed and native protein-DNA interfaces	128
24	Structures and on-target binding affinities of selected specific designs	129
25	Characterization of specificity at single base resolution	131
26	Detailed base contacts of designs evaluated by variant competition assay	133
27	HT-SELEX characterization of binder block designs	135
28	Proof-of-principle design of minor groove DNA binding proteins . . .	136
29	On-target inference-time search across the six core DNA targets . . .	137
30	Inference-time search under a combined on- and off-target reward . .	138
31	Distribution of interface minPAE among ordered designs, on-target binders and specific binders	139
32	How a position weight matrix becomes a training label	140
33	The two-dimensional block split of the two in-house panels	141
34	The re-ranking model over AlphaFold3 pair embeddings	142
35	Validation correlation over training for the DNA-only and multimodal models	143
36	Prospective evaluation on an orthogonality assay of 48 designs against 16 DNA targets	144

List of Tables

1	The sixteen DNA design targets	145
2	Design campaigns against DNA targets	146
3	Curated protein-DNA datasets	146
4	Composition of <code>dna_pwm</code> by source	147
5	Prospective evaluation on the 48-design orthogonality assay	147

Acknowledgments

I have so many people to thank for supporting me on this PhD journey. First and foremost, my family. My mother, father, and extended family, particularly my late grandmother, have all been immensely supportive and encouraging. My grandmother always hoped to see me finish my PhD, and although she unfortunately passed before she could, I'm glad that I was able to complete it and am grateful for the attendance of my great-aunt, her sister, at my defense in her stead.

Next my lovely partner, Milena Nikolajeva, who has kindly taken on the travel burden to come see me in Seattle during our long distance relationship. The desire to move to be with her helped maintain my motivation and focus during the inevitable peaks and valleys of the PhD experience.

Of course I must additionally thank all of my colleagues, mentors, coauthors, and friends in the lab and at the IPD. So many excellent scientists have come and gone over the years and I found it a privilege to speak to and learn from you all. From nanopores to DFT to immunology to artificial photosynthesis, the breadth and depth of understanding at this institution is truly special, and I will surely miss it. In particular, I make shout-outs to Pascal Sturmfels and Peik Lund-Anderson, whose close friendship both in and out of the lab has kept me sane and added greatly to the richness of my life; my collaborators on the DNA binding team, Yuliya Politanska, Enisha Sehgal, Raktim Mitra, Cameron Glasscock, Robert Pecoraro and Nayim Gonzalez; fellow members of the computational subgroup, Rohith Krishna, Andrew Kubaney, Jasper Butcher, Fanglei Xue, Jie Chen, Aiko Muraishi, Nate Corley, Rafi Brent, Max Kazman, Yanjing Li,

Tuscan Thompson, Woody Ahern, Brian Coventry, Joe Watson, David Juergens, Kieran Didi, Jacob Gershon, Sidney Lisanza, Andrew Favor, Jue Wang, and Doug Tischer; my 1820 Boylston building mates Preetham Venkatesh, Aditya Krishnakumar, and Jiji; and the many folks who went on hikes with me throughout the Cascades.

Outside of the lab I give thanks to my fellow MolES PhD cohort for their consistent, earnest and laughter-filled companionship, and my friends, both in Seattle and beyond. I am lucky to have so many cheering me on.

And finally, to my committee and to David Baker, for his mentorship and ideas, and for creating such a wonderful lab environment.

It has been a wonderful four years. Thank you.

1 Introduction

Nearly every process that reads, copies, repairs, or regulates the genome is carried out by a protein bound to DNA. Transcription factors control gene expression by occupying short sequences in promoters and enhancers, polymerases and helicases track along the helix to copy and unwind it, nucleases and recombinases and repair enzymes locate precise sites to cut, join, and mend, and histones and architectural proteins package the genome and set which regions are accessible [69, 83, 89]. These proteins span a wide range of selectivity. Histones and the processive polymerases and helicases bind DNA of essentially any sequence, while others must pick a single site out of a large excess of competing sequence, in the largest genomes discriminating a handful of base pairs against a background of billions.

Building sequence-specific DNA-binding proteins (DBPs) on demand would enable a broad set of applications. A protein that binds a chosen genomic address can serve as the targeting module for a base editor, a transcriptional activator or repressor, an epigenetic modifier, or a diagnostic sensor [54, 58, 97]. However, the tools that currently perform these functions, engineered zinc fingers, transcription activator-like effectors (TALEs), and CRISPR-Cas systems, are all descended from systems that nature evolved, and are therefore constrained by the scaffolds, folds, and recognition chemistries that evolution happened to explore [17, 49, 57]. With these tools we can reprogram and recombine functions that already exist, though reaching beyond that repertoire remains difficult. *De novo* protein design, the construction of proteins from first principles, offers a route past this limitation [44, 59]. My thesis is concerned with realizing that route for

DNA binding, and specifically with the computational methods, both generative and discriminative, needed to design proteins that recognize an arbitrary DNA sequence with high affinity and specificity.

This introduction provides the background for that goal. I first describe the chemistry of the two molecules involved, proteins and DNA, and the physical principles by which proteins fold and DNA forms a double helix. I then describe how proteins read DNA, that is, the structural and chemical basis of sequence-specific recognition, which is both the phenomenon we wish to reproduce and the source of its difficulty. I then cover the existing approaches to obtaining DNA-binding function, by discovering natural systems or by reprogramming them, which share a common ceiling set by their reliance on natural scaffolds. I then introduce *de novo* design and describe the work of Glasscock et al., the immediate predecessor to this thesis and the starting point for the methods developed in the following chapters. Finally, I describe the limitations of that work and state the research directions and goals of this thesis.

1.1 Biochemistry of Proteins and DNA

Proteins and nucleic acids are both linear polymers in which a one-dimensional sequence encodes a three-dimensional structure. A protein is a chain of amino acids joined by peptide bonds. The twenty canonical amino acids share a common backbone and differ in the side chain on the central α -carbon, which spans hydrophobic, aromatic, polar, and charged chemistries. This chemical diversity, arranged along the chain, allows proteins to fold into compact structures and carry out most of the active chemistry of the cell. A protein's fold is largely set by its sequence. Anfinsen's experiments established that

many proteins fold spontaneously and reproducibly to the conformation that minimizes their free energy — thus the native structure is encoded thermodynamically in the sequence [7].

DNA stores and replicates genetic information, and its structure is specialized for that role. It is a polymer of four nucleotides, adenine (A), thymine (T), guanine (G), and cytosine (C), each a base joined to a sugar and a phosphate. In Watson and Crick’s double helix, two antiparallel strands pair through complementary hydrogen bonding, A with T and G with C, with the sugar-phosphate backbones on the outside and the stacked bases within [99]. Under physiological conditions the helix adopts the right-handed B-form, with two grooves of different sizes winding along it, a wide, shallow major groove and a narrow, deep minor groove. This geometry is central to recognition, because the geometry and chemical properties of the atoms at the groove-facing edges are what allow a protein to sense which base pair sits at a given position. Additionally, the backbone carries a phosphate at every position and presents a uniformly negative, sequence-independent surface whose desolvation on binding is a major thermodynamic cost [79].

DNA is a flexible molecule, and the idealized B-form is only an average over an ensemble of conformations. Real DNA bends, twists, and varies its groove dimensions, and each sequence samples its own distribution of shapes, with some sequences favoring conformations that others adopt only rarely [66, 83, 84]. Additionally, these sequence-dependent shape preferences are themselves a recognition signal, both through general shape complementarity with the binding protein, and by modifying relative positions of

groove-facing atoms.

1.2 Protein-DNA Interactions in the Cell: Structure and Function

Proteins recognize DNA sequences through a small set of recurring physical mechanisms, combined in different proportions across different proteins [69, 83, 89]. These mechanisms include base readout, in which the protein senses the chemical identity of the bases directly, and shape readout, in which it senses the sequence-dependent conformation of the helix [83].

In base readout, a protein probes the hydrogen-bond donors, acceptors, and nonpolar groups exposed on the groove-facing edges of the base pairs (Figure 1). The major groove is more informative than the minor groove. Each of the four base pairs (A:T, T:A, G:C, C:G) presents a distinct pattern of contacts along its major-groove edge, while in the minor groove A:T and T:A present nearly identical edges, as do G:C and C:G. A single hydrogen bond leaves a base pair ambiguous, since small side-chain movements create degeneracies, but a pair of hydrogen bonds made simultaneously to the same base, a bidentate contact, fixes the geometry tightly enough to confer specificity. Across solved complexes there are clear, family-independent amino acid-base preferences, with arginine and lysine reading guanine, asparagine and glutamine reading adenine, and threonine reading thymine, and with arginine-guanine as the most common bidentate motif [68]. Arginine dominates because its guanidinium group donates two hydrogen bonds at once to the O6 and N7 atoms of guanine, and the contact to O6 distinguishes guanine from adenine, which shares the N7. Base recognition typically involves networks

of hydrogen bonds spanning several base steps, so specificity is a collective property of the interface [83].

Shape readout accounts for specificities that arise from the DNA's local geometry, complementing base readout. Because the sequence of a site influences the local conformation of the helix, a protein can recognize a sequence indirectly, by recognizing the shape that sequence adopts [83, 84]. This occurs at both global and local scales. Globally, a binding site may be intrinsically bent or driven into an A-form or Z-form conformation, and the protein selects for that deformability; the TATA-box-binding protein, for example, inserts a β -sheet into a sharply unwound and bent minor groove [56]. Locally, the width of the minor groove varies with sequence. A-tracts narrow it, and a narrow minor groove focuses the negative electrostatic potential of the backbone into a pocket that basic side chains, most often arginine, recognize through enhanced electrostatic attraction [83, 102]. Most of the interface engages this sequence-independent backbone, with roughly two-thirds of all contacts made there, contributing binding energy and positioning; a typical interface buries roughly two dozen residues against about a dozen nucleotides [83]. Additionally, recognition can involve water-mediated hydrogen bonds and induced-fit rearrangements in which the protein, the DNA, or both distort only upon binding [55].

Proteins deploy these mechanisms from a limited repertoire of structural scaffolds. DNA-binding domains fall into more than seventy structural superfamilies, though a few motifs dominate (Figure 2) [83]. The helix-turn-helix motif, found in the classic bacterial repressors (λ , Trp, Lac) and in eukaryotic homeodomains, inserts a recognition helix into

the major groove, where its side chains read the base edges [40, 72, 80]. Cys₂His₂ zinc fingers use a compact $\beta\beta\alpha$ fold stabilized by a zinc ion, and their short α -helix contacts roughly three base pairs. Because these fingers occur in tandem arrays, a protein can read a longer site by stringing several together, and this modularity made zinc fingers the first workhorse of DNA-binding-protein engineering [30, 57, 105]. Additionally, basic leucine-zipper and helix-loop-helix proteins dimerize and grip the major groove from opposite faces, and minor-groove binders such as the HMG-box and TBP rely predominantly on shape.

Taken together, these mechanisms show that recognition is combinatorial and context-dependent [68, 86, 89]. The favored contacts (Arg-G, Asn/Gln-A) are real and statistically robust, but they operate within interfaces whose specificity depends on the combined effect of base readout, shape readout, backbone contacts, water, and conformational flexibility. Thus, closely related transcription factors with nearly identical binding domains can select different sites, and functionally distinct factors can prefer the same site [83]. Furthermore, recognition in the cell is more complicated still, as a typical transcription factor occupies thousands to tens of thousands of genomic sites, and its specificity *in vivo* depends as much on cooperative binding partners, local concentration, and chromatin accessibility as on the intrinsic preference of its binding domain [89].

1.3 Existing Discovery Approaches to Protein-DNA Interactions

Because first-principles prediction of DNA-binding specificity remains an open problem, sequence-specific DNA-binding functions have in practice been obtained by two routes:

discovering the proteins that nature has already evolved, and reprogramming a small number of natural scaffolds to new targets (Figure 3).

The first route is discovery and characterization. Nature has evolved many DNA-binding proteins, and much of molecular biology and biotechnology has proceeded by finding, characterizing, and repurposing them. Currently, high-throughput methods map the specificities of natural transcription factors at scale, and SELEX, protein-binding microarrays, and chromatin immunoprecipitation assays generate the binding profiles collected in databases such as JASPAR [15, 20]. Metagenomic sequencing has enabled large-scale computational discovery methods, as Altae-Tran and colleagues, for example, mined billions of genomic fragments with a scalable clustering algorithm to uncover the functional diversity of rare CRISPR-Cas systems, expanding the catalog of natural programmable nucleases available for engineering [4]. The CRISPR-Cas systems found this way, from the Cas9 nuclease to compact effectors such as Cas14, have become the dominant tools for targeting DNA, because they are programmable, with their specificity set by a guide RNA [39, 49, 97]. However, they retain natively-evolved limitations: they are large, they require an RNA component, and they are constrained by protospacer-adjacent-motif requirements that limit which sites can be addressed.

The second route is reprogramming, taking a natural DNA-binding scaffold and altering its specificity. Cys₂His₂ zinc fingers, with their modular three-base-pair fingers, were the first scaffold engineered this way, through selection and rational design, although assembling fingers into arrays that behave predictably has proven laborious because of context-dependence between adjacent fingers [48, 57, 105]. Alternatively, TAL

effectors offer a cleaner, nearly modular code of one repeat per base pair, and became widely used for genome targeting, at the cost of large, highly repetitive genes that are difficult to synthesize and deliver [17, 52]. Homing endonucleases were reprogrammed computationally, and the Baker and Stoddard laboratories used Rosetta to redesign the binding and cleavage specificity of enzymes such as I-CreI and I-MsoI at individual base positions, an early demonstration that protein-DNA specificity could be engineered *in silico* [10, 11, 22, 91, 93]. Additionally, structure-based methods were developed to predict specificity, for example for C2H2 zinc fingers [108].

These approaches share a limitation that motivates this thesis. Each of them begins from a fixed natural scaffold, with a particular fold, a particular orientation relative to the DNA, and a particular repertoire of recognition chemistries. The scaffold thereby constrains which target sequences can be reached and how precisely the protein can be tailored to them [37].

1.4 The Promise of De Novo Design

De novo protein design breaks this dependence on natural scaffolds by building proteins from scratch. One specifies a target structure or function and constructs a new sequence, and often a new fold, to satisfy it. The proof of principle was established with the design of Top7, a protein with an entirely novel fold, designed with the Rosetta energy function and validated by crystallography [62]. Successive work iterated on these physics-based methods, and then more recently, deep learning has accelerated development of this field and broadened its scope [59], which I discuss in detail in Chapter 2.

De novo design is well suited to DNA binding in principle. A design approach can choose its scaffold size and topology freely, independent of the folds and recognition chemistries that natural transcription factors provide, and it can in principle target an arbitrary sequence with a compact, single-chain, easily delivered protein and couple it to effector functions of our choosing [37]. However, DNA is a demanding target for *de novo* design. Its recognizable surface is highly polar, so binding requires paying a large desolvation cost and satisfying the hydrogen-bonding requirements of both the phosphate backbone and the base edges, the four bases differ only by subtle changes in atom placement, and specificity depends on long, flexible polar side chains placed with near-atomic precision [37, 79].

1.5 Prior Work: Glasscock et al.

The immediate predecessor to this thesis is the work of Glasscock et al., which demonstrated the first computational *de novo* design of sequence-specific DNA-binding proteins, built from scratch to recognize chosen target sequences through base contacts in the major groove [37].

The design strategy adapted the *de novo* binder-design paradigm to the demands of DNA (Figure 4). To provide shape complementarity to the double helix and satisfy the phosphate backbone, the authors assembled a custom library of roughly 26,000 small (<65-residue) helix-turn-helix scaffolds, mined from metagenomic sequence data and filtered by AlphaFold2 confidence. Base-specific contacts were seeded with an extension of the RIFdock method [19]. For each DNA target they enumerated a Rotamer Interaction Field, a large set of disembodied side-chain rotamers making favorable, statistically

frequent bidentate hydrogen bonds to the exposed base atoms in the major groove, and docked scaffolds against it to find placements that satisfied both backbone-phosphate hydrogen bonds and base-contacting rotamers. Sequences were then designed onto these docks with Rosetta or with LigandMPNN, the DNA- and ligand-aware extension of ProteinMPNN [26], and the designs were filtered on binding energy, interface hydrogen bonding, side-chain preorganization, and agreement between the AlphaFold2 prediction and the design model.

Screened by yeast-surface display with fluorescently labeled DNA targets and deep sequencing, 97 designs were strongly enriched for their intended target. Of these, 44 showed detectable binding when tested individually, and in an all-by-all screen of those 44 against thirteen DNA targets, 14 bound with specificity closely matching their design models, giving validated sequence-specific binders for five of the thirteen targets with affinities in the tens to hundreds of nanomolar (Figure 5). A crystal structure of one designed complex matched the design model closely, and the designs functioned as transcriptional regulators in both *E. coli* and mammalian cells. This work established that sequence-specific DNA recognition can be designed *de novo*, and that the targetable sequence space extends beyond what natural proteins bind.

1.6 Research Directions and Goals

These results also exposed the bottleneck that this thesis sets out to address. The translation from computational design to validated binder was very inefficient, on the order of a few dozen binders from more than a hundred thousand ordered designs, which makes each campaign costly. Additionally, many of the binders that did emerge

recognized their target with poor specificity, and some target sequences resisted every design attempt. Furthermore, the *in silico* filters, primarily AlphaFold2 refolding metrics, were only weakly predictive of experimental success. The generative step could propose plausible DNA-binding proteins, but two advances were still needed: generative models expressive enough to produce high-quality, base-reading interfaces at scale, and discriminative models accurate enough to predict, before synthesis, which designs will bind and which sequence they will prefer. Improving generation and prediction to close these two gaps is the subject of this thesis.

2 Generative Models for DNA Binder Design

Modern protein design pipelines broadly involve two stages: a generative stage that proposes candidate proteins, and a discriminative stage that predicts which candidates will bind. This chapter concerns the generative stage, whose task is to sample protein structures and sequences that fold, that place side-chain atoms in hydrogen-bonding geometry against a DNA target, and that are diverse enough to supply many independent candidates per target. Denoising diffusion models are used for this task, built by fine-tuning or retraining the neural networks developed for protein structure prediction. Section 2.1 develops the diffusion framework. Section 2.2 describes the representations and network architectures used to model protein and DNA structure. Section 2.3 combines them in the RFdiffusion family of generative models and applies them to DNA binder design, covering RFdiffusion, CA-RFdiffusion, and RFdiffusion All-Atom, and the limitations of each that motivated RFdiffusion3 (Chapter 3).

Generative modeling of proteins has been approached with every major class of generative model: autoregressive language models trained on sequences [32, 70], variational autoencoders over structural ensembles [71], and graph-based models of the sequence-structure map [46]. For the generation of three-dimensional backbones, diffusion models have been the most successful, and are the subject of this chapter.

2.1 Diffusive Generative Models

A generative model seeks to draw new samples from a data distribution $p(x)$ given only a finite set of examples. Denoising diffusion probabilistic models (DDPMs) do this by

defining a fixed process that gradually destroys data by adding noise, and training a neural network to reverse it (Figure 6) [42]. A network that can reverse a single noising step can, applied repeatedly from an initial sample of pure noise, produce a sample from the data distribution.

The Forward (Noising) Process

Let x_0 be a data point, for us ultimately a protein structure. The forward process is a fixed Markov chain that adds a small amount of Gaussian noise at each of T steps, according to a variance schedule $\beta_1, \dots, \beta_T$:

$$q(x_t | x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I}\right).$$

Because the noise is Gaussian, the chain has a closed form: defining the signal rate $\alpha_t = 1 - \beta_t$ and its cumulative product $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, the noised sample at any step t can be written directly in terms of the clean data,

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}).$$

The schedule is chosen such that $\bar{\alpha}_T \approx 0$, at which point the distribution of x_T is indistinguishable from $\mathcal{N}(0, \mathbf{I})$. Generation therefore starts by drawing $x_T \sim \mathcal{N}(0, \mathbf{I})$.

The Reverse (Denoising) Process

Generation runs the chain backward, sampling from $p_\theta(x_{t-1} | x_t)$. For small enough steps this reverse conditional is itself approximately Gaussian, and a neural network with parameters θ is trained to predict its mean. In the parameterization of Ho et al.,

the network $\epsilon_\theta(x_t, t)$ is trained to predict the noise ϵ that was added, which reduces the training objective to a simple denoising regression,

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0, \epsilon, t} \left[\left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t \right) \right\|^2 \right],$$

with t sampled uniformly over the T steps [42]. The network can equivalently be trained to predict the clean data x_0 directly, and the two are related by the closed form above. For protein structures the network is trained to predict x_0 , so its output at every step is a fully formed structure. Generation then alternates a denoised prediction with a small re-noising step, which makes the trajectory follow the reverse of the forward chain and converge on the data manifold.

Conditional Generation

Design requires constraining the generated structure to a specification. In conditional generation the denoising network receives a condition c alongside the noised structure x_t and the timestep t , and is trained to model $p_\theta(x \mid c)$. During training the condition is annotated from the native structure of each example. At inference it is set by the designer to describe the target protein. The condition is supplied at every step of the reverse process, so it constrains the entire denoising trajectory.

Conditions are supplied in two forms. The first is structural, where the coordinates of a fixed functional motif, a binding partner, or a target surface are held fixed in the representation while the surrounding protein is generated, so the network builds a structure in their geometry [100]. The second is an auxiliary feature channel aligned to

the tokens or atoms, a one-dimensional feature vector of length equal to the number of residues or the number of atoms that labels each element with a property the design should satisfy. This thesis uses hydrogen-bond conditioning, in which the feature marks which target atoms are hydrogen-bond donors or acceptors for the designed protein to satisfy. An all-atom network reads this per-atom label directly and places complementary side chains at the interface [18]. Solvent accessibility, specified as a per-atom burial label, and the position of the protein’s center of mass relative to the target are supplied the same way.

Conditions are annotated from native structures and supplied on a fraction of training examples, with the remainder trained unconditionally, which teaches a single network both unconditional sampling and adherence to any subset of the conditions. Adherence to a supplied condition can be sharpened at sampling time by classifier-free guidance, in which each denoising step averages a conditioned and an unconditioned forward pass of the same network and extrapolates between them,

$$\tilde{\epsilon}_{\theta}(x_t, c) = (1 + w) \epsilon_{\theta}(x_t, c) - w \epsilon_{\theta}(x_t),$$

with w setting how strongly the sample is pulled toward the condition [43].

In protein design the condition may be a topology, a symmetry, a functional motif, or a DNA target. How the DNA-binder generative model is conditioned on a DNA target, and on the specific interface hydrogen bonds that encode sequence-specific readout, is developed in Section 2.3 and Chapter 3.

Why Diffusion

Diffusion models replaced earlier generative approaches to protein design for several reasons. They train stably, by a regression objective. They produce diverse samples spread across the data distribution, which a design campaign requires when it needs thousands of independent candidates against one target. On image benchmarks they surpass the best adversarial models in sample quality [27]. The reverse process is iterative, so conditioning information is applied at every step and a single trained model can be directed at many design tasks. Practical variants improve efficiency, for example by running the diffusion in a compressed latent space [85].

2.2 Modeling of Proteins and DNA

In order to apply diffusion to protein structure, we require a representation of the protein, the nucleic acid, and the atoms of the interface between them that a neural network can operate on. The representations used for protein design are largely inherited from structure prediction. This section describes the main options — residue frames, full-atom representations, and learned latents (Figure 7) — and the network architectures and attention patterns that accompany them.

Every representation of three-dimensional structure is subject to one constraint. A protein structure is a set of coordinates, and the same structure translated or rotated in space is physically identical, so a model of structure should be equivariant to the group of rigid motions, $SE(3)$. One route to this is architectural, where the network is constructed such that rotating its input rotates its output identically, and it is trained

against a loss that is invariant to the global pose in which a structure is expressed [34, 50]. A second route is data augmentation, in which a random rigid rotation of each training example is sampled during training and the network learns approximate equivariance from the data. The all-atom models described below take this route [1, 18].

Residue Frames, Backbone Atoms, and Tokenization

The most compact representation of protein structure treats each residue as a single token whose geometry is a rigid frame. AlphaFold2 introduced the frame definition used by the structure-prediction and design models that followed: each residue is a frame consisting of a translation, the position of the C_α atom, and a rotation, the orientation of the N- C_α -C backbone triad [53]. A nucleotide is represented the same way, RoseTTAFoldNA placing a frame at each base from the phosphate coordinate and the OP1-P-OP2 orientation [13]. The remaining backbone atoms follow from the frame by ideal geometry, and each residue’s side-chain atoms are placed by the torsion angles about its rotatable bonds. A structure is then a set of frames, one per token, and the network reasons at the granularity of tokens, which keeps the representation small: one token per residue, in place of a node for each of its heavy atoms.

Although this coarse-grained representation is efficient, it limits what the model can express at the interface. Sequence recognition depends on the positions of the side-chain and base-edge atoms that form the interface (Chapter 1), and the frame representation encodes those positions indirectly, through the backbone frame and its torsion angles. Most of the model’s computation is at the frame level: attention layers operate on residues, and torsion angles are predicted by a head auxiliary to the residue-to-residue

attention. The atomic geometry of the interface is therefore learned implicitly.

Full-Atom and Latent Representations

A more expressive option represents structure at the level of individual atoms. RoseTTA-Fold All-Atom keeps the frame-and-torsion representation for biopolymers but adds an atom-bond graph for arbitrary small molecules and for covalently modified or atomized residues, presenting an assembly to the network as a set of residue frames together with individually represented atom nodes [60]. AlphaFold3, and its open reimplementations Chai-1 and Boltz, dispense with frames altogether: every atom is represented explicitly, and all-atom coordinates are denoised directly by a diffusion module [1, 21, 104]. An all-atom representation places the atoms of a side-chain-base hydrogen bond explicitly, and requires the network to attend over every atom rather than every residue, which motivates the sparse and local attention patterns described below. These predictors are used as discriminative filters for protein-DNA binding in later chapters.

A third option represents the atomic detail in a learned latent. La-Proteina represents the backbone explicitly while capturing the sequence and all remaining atomistic detail in per-residue latent variables of fixed size, and uses flow matching in this partially latent space to model the joint distribution over sequence and full-atom structure [36]. Because the latent has a fixed dimension for every residue, side chains of any amino-acid identity are represented by the same number of variables, and the model scales to proteins of many hundreds of residues. The same latent approach has since been extended to fully atomistic binder design [28].

Network Architectures and Attention Patterns

The attention pattern of a network follows from its representation. Structure prediction models process an evolutionary multiple-sequence alignment together with a pairwise representation of residue-residue relationships, then build coordinates in a structure module. AlphaFold2 introduced Invariant Point Attention, which operates on the residue frames in an $SE(3)$ -invariant way, and trains against the Frame-Aligned Point Error, which measures atomic error in every residue’s local frame and is by construction invariant to global pose [53]. RoseTTAFold reaches comparable accuracy with a three-track architecture that passes information between a one-dimensional sequence track, a two-dimensional pairwise track, and a three-dimensional coordinate track, refining all three together [12], building on earlier work that predicted inter-residue geometries [88, 107].

An all-atom representation needs a different attention pattern, because attending all-to-all over every atom in a complex is prohibitively expensive on current GPU hardware. AlphaFold3 keeps a coarse token-level transformer for long-range reasoning and couples it to a fine atom-level module, moving information between the two by cross-attention and restricting atom-level attention to spatially local neighborhoods [1]. This token-atom coupling, with local attention at the atomic scale and cross-attention pooling between scales, is the pattern the all-atom design model in Chapter 3 adopts as well.

In order to reduce the cost of generating a multiple-sequence alignment, protein-language-model based approaches replace it with learned residue embeddings. For example, ESMFold and OmegaFold predict structure from a single sequence, trading some accuracy

for speed [64, 106].

These representations and the networks that operate on them were developed for structure prediction. The same networks are used for design, as a network trained to map an explicitly represented sequence to a folded structure is retrained as the denoiser of a diffusion model, mapping an input context and a corrupted structure to a generated design.

2.3 Diffusion Models for Protein Design

The RFdiffusion family applies the diffusion framework of Section 2.1 to the structural representations of Section 2.2, using a structure-prediction network as the denoiser, either fine-tuned from pretrained weights or trained from scratch [100]. Earlier deep-learning approaches to protein design were limited by backbone quality: network hallucination optimized sequences to be confidently predicted by a structure network [8]; single-step inpainting scaffolded motifs deterministically, giving one solution per input [98]; and early protein diffusion models generated backbones that folded poorly [5, 109]. This section describes RFdiffusion, then two adaptations of it to DNA binder design, CA-RFdiffusion and RFdiffusion All-Atom, which represent the protein-DNA interface in increasing atomic detail.

RFdiffusion

RFdiffusion fine-tunes RoseTTAFold into the denoising network of a diffusion model operating on the residue-frame representation of Section 2.2 [100]. The forward process noises a structure by perturbing each residue’s C_α translation with three-dimensional

Gaussian noise and its frame orientation by Brownian motion on the manifold of rotations, such that after enough steps the frames are random. At each reverse step the network predicts the denoised structure directly, a fully formed backbone, and a small noised step toward that prediction becomes the input to the next iteration, as in Section 2.1. Three choices proved important: a mean-squared-error loss on the predicted frames, which depends on the global reference frame and so keeps the coordinate frame consistent from one step to the next; SE(3)-equivariance inherited from RoseTTAFold, so the process behaves correctly under global rotation; and self-conditioning, letting the network condition on its own previous prediction, analogous to recycling in AlphaFold2. Fine-tuning from pretrained RoseTTAFold weights outperformed training from untrained weights for an equivalent length of time. Once a backbone is generated, its sequence is designed separately with ProteinMPNN [25], and designs are validated *in silico* by checking that AlphaFold2, run on the designed sequence alone, re-predicts the intended structure with high confidence.

Under different conditioning, a single RFDiffusion model addresses a wide range of design problems [100]: unconditional generation of diverse, designable monomers far from anything in the training set; fold and topology conditioning toward a desired architecture such as a TIM barrel or an NTF2 fold; symmetric assembly, where the equivariance lets a single subunit be replicated under a chosen point group to build oligomers of essentially any symmetry [103]; functional-motif scaffolding, holding the fixed coordinates of an active site, binding loop, or epitope in place while building a supporting scaffold; and binder design, generating a protein that folds against and binds a specified target surface [19]. The same paradigm has been extended to joint

sequence-structure generation, functional and multistate design, and programmable generative models over protein space [38, 47, 65, 96]. The design of a protein that binds a specific DNA sequence is a binder-design problem in which the interface must place side-chain atoms against the DNA base edges precisely enough to discriminate between sequences.

Extending RFdiffusion to DNA Binder Design

RFdiffusion models the protein backbone as a set of frames. To enable binder design to DNA targets, we must represent the DNA target in the generative model alongside the protein. My first approach did so while keeping the frame representation, adapting CA-RFdiffusion, the framework of Lauko et al. originally developed for enzyme design [63, 87].

CA-RFdiffusion is a two-model scheme. A diffusion model generates the alpha-carbon coordinates of a backbone, a C_α trace, and a separate refinement model recovers the full N-C α -C frames from that trace in a single step; both are fine-tuned from the same pretrained structure-prediction network [63]. The diffusion model is trained with a loss on alpha-carbon positions alone, which leaves it responsible for placing the C_α trace, and the frame orientations are recovered by the refinement model.

To reach DNA, I fine-tuned the CA-RFdiffusion diffusion model from RoseTTAFoldNA, the protein-nucleic-acid structure prediction model, such that the model natively represents the DNA alongside the protein [13, 87]. During training, we sampled native protein-DNA complexes from the Protein Data Bank 60% of the time, monomeric proteins from the Protein Data Bank 30% of the time for fold diversity, and predicted

protein-DNA complexes from a distillation set 10% of the time, each example cropped tightly around a protein contact with a DNA base or backbone atom to focus the model on the geometry of recognition. A contact loss was added to improve the placement of the recognition helix against the DNA. At inference, designs were generated against DNA target structures predicted by RoseTTAFoldNA rather than against idealized B-form DNA, since the target structure then includes the sequence-dependent deviations from ideal geometry discussed in Chapter 1. Sequences were designed onto the generated backbones with the nucleic-acid-aware LigandMPNN [26].

***In Silico* Filtering with RoseTTAFoldNA**

Designs were filtered by refolding each designed sequence in the presence of its DNA target with RoseTTAFoldNA and scoring the prediction against the design model [13]. This filter was chosen on retrospective evidence. Across roughly 500 designs from earlier campaigns that had all been tested experimentally, of which 11 bound specifically and 26 bound without specificity, requiring a predicted probability of binding above 0.5, a protein-only C_α RMSD below 2 Å, and a DNA-aligned protein RMSD below 5 Å raised the fraction of binders in the retained set from roughly 7% to roughly half. Because designs would be selected on this criterion, the generative model was developed to produce backbones that satisfy it.

Successive fine-tuning checkpoints were compared on the fraction of generated backbones, and of designed sequences, that pass these cutoffs. The fraction of backbones passing rose from roughly 6% for the best previous model to roughly 15%, and the fraction of sequences passing rose from roughly 1.0% to roughly 2.7%, with the best performance

at 14 epochs. However, both fractions declined at 16 epochs. Overtraining of this kind was more severe for the earlier full-frame diffusion model and was only partly mitigated by restricting the diffusion to alpha-carbon positions.

The design campaign ran in two rounds (Figure 8). In the first, 50,000 backbones were generated by free diffusion against the target, 12 sequences were designed on each with LigandMPNN, and the resulting complexes were refolded with RoseTTAFoldNA, of which 3,667 designs (0.78%) passed. Backbones that passed were resampled by partial diffusion, in which a design is renoised part way along the forward trajectory and denoised again, and the resampled backbones were redesigned and refiltered, of which 47,037 designs (11.6%) passed. Rosetta interface metrics were applied on top of the RoseTTAFoldNA cutoffs, and approximately 10,000 designs per target were ordered.

Wet-Lab Results

Approximately 60,000 designs against 6 DNA targets were ordered as a synthetic oligonucleotide library, displayed on the surface of yeast, and screened for binding to biotinylated DNA targets by fluorescence-activated cell sorting, following the assay of Glasscock et al. [37]. We sorted at successively lower target concentrations, from 1 μ M down to 150 pM. This gave a concentration-dependent decrease in the gated population, which remained above the no-DNA control throughout (Figure 9). Sequencing of the sorted populations identified 2,629 designs (4.4%) enriched against at least one target at 1 μ M avid concentration, of which only 17 (0.031%) were enriched against their intended target. Unfortunately, designs were as likely to be enriched against any individual off-target as against the target they were designed for.

The comparison campaign, in which backbones were generated by RIFdock as in Glasscock et al., ordered approximately 100,000 designs against 12 DNA targets, a subset of the 133,761 designs of Section 1.5 [37]. Sequencing of the sorted populations identified 320 designs (0.32%) enriched against at least one target in the panel, and 27 (0.027%) enriched against their intended target and specific for it. Therefore, diffusion raised the rate at which a design bound DNA by roughly an order of magnitude, from 0.32% to 4.4%, and gave the same rate of on-target specific binding, 0.031% against 0.027%.

However, a small number of designs achieved both high affinity and high specificity. Eight designs against a single target bound it strongly while showing little binding across a panel of twelve other DNA targets, and the tightest of these bound with a dissociation constant of 26.9 ± 7.0 nM measured by biolayer interferometry (Figure 10).

The RoseTTAFoldNA metrics used to select these designs anticorrelated with the experimental outcomes (Figure 11). We compared nonbinders, nonspecific binders, and specific binders, and the specific binders scored worse than the nonbinders on all four metrics: higher interface predicted aligned error, higher ligand RMSD, lower predicted probability of binding, and lower pLDDT. Unfortunately, the filter that had enriched binders among RIFdock designs was uninformative on designs from the diffusion model.

These results indicate that the deficiency of the design model is in generating high-quality interfaces. The order-of-magnitude increase in binding rate shows that the diffusion model generated backbones better suited to the DNA surface than those found by docking a fixed scaffold library, while the flat rate of specific binding shows that the

hydrogen bonds between protein side-chain atoms and DNA base atoms were of lower quality. The diffusion model operates on alpha-carbon coordinates and on DNA frames placed at the phosphate group. The side-chain atoms that read the bases are placed by LigandMPNN after generation is complete, and the base atoms themselves reach the model only through the phosphate frame, so the backbone is positioned against the DNA before any sequence-specific contact exists to constrain it. RIFdock seeds each dock with side chains already making bidentate hydrogen bonds to base atoms, so in that pipeline the base contacts are fixed first and the scaffold is placed to accommodate them.

RFdiffusion All-Atom

In RFdiffusion and CA-RFdiffusion the diffusion process places backbone atoms alone, and the side chains that contact the DNA bases are added afterward by sequence design. During generation the model therefore has access only to the protein backbone and the DNA backbone frames, and must infer from that geometry which backbones will support a sequence-specific interface once side chains are designed. RFdiffusion All-Atom represents the contacting atoms in the generative model itself. It is built on RoseTTAFold All-Atom, whose representation keeps residue frames for the bulk of the protein and adds an explicit atom-bond graph for ligands and for atomized residues (Section 2.2) [60], and is fine-tuned from RoseTTAFold All-Atom weights into a diffusion denoiser (Figure 12). It was applied first to enzyme and small-molecule design, where the atoms of a catalytic site or bound ligand define the motif and the network generates a protein around them.

A residue that must be reasoned about atom by atom is atomized, and is represented as the explicit atoms of its side chain, nodes in the atom-bond graph with atom, bond, and chirality features, tied back into the chain by atom-to-residue bonds, so a single structure holds most residues as frames and a chosen few as atoms [60]. Guideposting then extends atomization, as the functional atoms of a motif are supplied as atomized nodes that the network must satisfy, and the network assigns each of them to a residue position of its own choosing and builds the rest of that side chain, so the motif is specified without fixing its position in the sequence. In tip-atom scaffolding the guidepost atoms are the side-chain tips that make a contact, and the network places a residue that reproduces them. The combination of atomic and frame representations is partial, since only the guidepost atoms are represented explicitly, the remainder of the protein is represented as frames, and the two are joined by the atom-to-residue bonds.

Extending RFdiffusion All-Atom to DNA Binder Design

For DNA, I adapted the tip-atom motif-scaffolding framework of RoseTTAFold All-Atom to generate binders [60, 87]. The model was fine-tuned from RoseTTAFold All-Atom weights on the generative task, using native protein-nucleic-acid complexes from the Protein Data Bank together with a distillation set of predicted transcription-factor structures from RoseTTAFoldNA, all cropped around a protein contact with a DNA base or backbone atom [13]. The model was trained on a tip-atom placement task and a motif scaffolding task. In tip-atom placement, the side chains that contact DNA are reduced to their contacting atoms, and the model is given only those tip-atom positions and must recover the rest of each side chain. In motif scaffolding, the model is given

the backbone coordinates of a contiguous stretch of protein near the DNA and must build the remaining scaffold. Radius-of-gyration, percent-helicity, and center-of-mass conditioning were added for finer control over the generated backbone [87].

Several changes to the representation and to the training process accompanied the move to an all-atom model. The nucleotide frame was moved from the phosphate group to the sugar, placing the frame origin closer to the base and to the glycosidic bond that orients it. DNA residues contacted by the protein were atomized alongside the contacting protein side chains, such that the atoms on both sides of an interface hydrogen bond are represented as explicit nodes in the sequence and pair tracks. Guideposting was applied at the residue level and at the atom level, such that a contact can be specified either as a whole residue whose position in the chain is left free or as individual atoms the model must satisfy. The training loss was upweighted on the atoms at the interface. Frame translations and rotations were generated by flow matching, which trains a network to predict a time-dependent velocity field transporting a simple prior distribution to the data distribution and integrates that field at inference to produce a sample, in place of the discrete denoising chain of Section 2.1. The position of the generated protein relative to the target was set by centering the initial noise on the center of mass of the diffused region, specified by an ori token, the mechanism Chapter 3 describes in full.

At generation time, the tip atoms that seed a design are the side-chain atoms of amino acids sampled to hydrogen-bond the DNA base edges in the major groove, placed by RIFGen, a rotamer-interaction-field method [19]. The all-atom model then builds a protein that holds those seeded contacts against the target, so the designed interface is

built around a set of base contacts specified at the atomic level (Figure 13).

***In Silico* Filtering with AlphaFold3**

By the time this model was applied, AlphaFold3 and its open reimplementations had been released, and a second retrospective analysis over the accumulated experimental data compared them against RoseTTAFoldNA as filters [1, 13, 104]. We ranked designs by each metric individually and measured the area under the receiver operating characteristic curve against experimental binding (Figure 14). AlphaFold3 reached 0.63 on interface predicted TM score, 0.61 on predicted TM score, and 0.59 on the deviation between the prediction and the design model; Boltz reached 0.58 to 0.59 on the same three metrics; RoseTTAFoldNA reached 0.57 on deviation from the design model, 0.54 on interface predicted aligned error, and 0.50 on the predicted probability of binding, which is chance. However, the discriminative power of every one of these metrics is modest, and AlphaFold3 appears to be informative mainly in the extreme tail. If we take the top 1% of the retrospective set by DNA-aligned protein deviation from the design model, the experimental success rate rises from roughly 0.08% to roughly 0.23%. The most discriminative AlphaFold3 metric is therefore this geometric agreement rather than any confidence metric. On this evidence designs were filtered with AlphaFold3 in place of RoseTTAFoldNA.

The Design Campaign

The campaign generated an interface first and a scaffold around it. DNA targets were prepared both as conformational ensembles predicted by Chai-1 and as DNA conformations taken from crystal structures [21]. Interfaces were seeded by sampling

hydrogen-bonding side chains over the major groove, arginine, asparagine and glutamine over guanine and adenine bases, at 10,000 seed combinations per DNA target. Tip-atom diffusion generated three helices per seed combination, 30,000 per target, which were filtered on Rosetta total energy below 300 and helicity above 75%, given sequences by LigandMPNN, and filtered again on the LigandMPNN base score [3, 26]. Motif-scaffold diffusion then built 50 complete backbones per surviving helix, 25,000 per target, and LigandMPNN designed sequences onto those. Each sequence was predicted in complex with its DNA target by AlphaFold3 and required to fall within 2 Å protein C $_{\alpha}$ RMSD and 5 Å DNA-aligned RMSD of the design model, with interface predicted TM score above 0.75. Passing designs were resampled by partial diffusion and refiltered.

Generation was split into a helix step and a scaffold step because the two steps fail at different rates. Many arrangements of seeds admit no single helix that can hold all of them at once, so most seed arrangements fail at the helix step. However, once a seed arrangement and a helix that satisfies it exist, many scaffolds can be built around that helix. Therefore, we filtered hard at the first step, which removed the arrangements that cannot work before any scaffolding compute is spent on them.

Wet-Lab Results

26,000 designs against 19 DNA targets were ordered as a synthetic oligonucleotide library, displayed on the surface of yeast, and screened for binding to biotinylated DNA targets by fluorescence-activated cell sorting, following the assay of Glasscock et al. [37]. Sequencing of the sorted populations identified designs enriched against at least one target at 1 μ M avid concentration at a rate of 4.6%. 192 of the enriched designs,

spanning all 19 targets, were ordered as individual genes and characterized in a 96-well format. 20 of these bound their intended target, spanning 13 of the 19 targets. We tested those 20 against the full panel of DNA targets. We counted a design that bound only its intended target as sequence specific, and a design that bound its intended target together with a minority of the panel as selective. This gave 4 sequence-specific binders and 3 selective binders, 0.027% of the ordered library (Figure 15).

CA-RFdiffusion gave 4.4% enrichment against any target and 0.031% on-target specific binders, and the RIFdock pipeline of Glasscock et al. gave 0.32% and 0.027% [37]. Therefore, the all-atom model reproduced the binding rate of the alpha-carbon model, and gave the same rate of sequence-specific binding as the docking pipeline. In total, three generations of generative model and three *in silico* filters have given the same rate of sequence-specific binding.

Two features of the campaign account for the flat specificity rate, and both concern the filter rather than the generative model. The pipeline optimized AlphaFold3 metrics directly by resampling designs with partial diffusion and redesigning them until they passed the cutoffs. However, optimization against a filter can reach regions of sequence space the filter has not learned well, where it pathologically reports good metrics, at which point it stops being informative. Every sequence in the order passed the AlphaFold3 cutoffs, and the rate of binding did not improve over the previous campaign.

Additionally, AlphaFold3 does not separate the intended target from the rest of the panel. We folded the binders recovered from the sort against every DNA target in the set, which gives interface predicted TM scores between 0.6 and 1.0 for nearly every

design-target pair, with the intended target scoring no higher than the others. The measured half-maximal binding concentrations over the same panel vary widely between design-target pairs by comparison: some designs bind one target and nothing else, others bind half the panel, and some bind an off-target more strongly than the target they were designed against. AlphaFold3 assigns a high interface predicted TM score to nearly every design-target pairing, independent of the target sequence. Therefore, selecting for specificity requires a model trained to predict it, which is the subject of Chapter 4.

Toward RFdiffusion3

Both of the above models are frame-based. CA-RFdiffusion generates C_α positions and recovers frames by refinement, and RFdiffusion All-Atom represents a chosen subset of residues as explicit atoms against a frame scaffold. In both, most interface atoms are represented through the backbone frames and their torsion angles, even though protein-DNA recognition depends on the explicit positions of those atoms.

Both campaigns raised the rate at which a design binds DNA by an order of magnitude over docking, 4.4% and 4.6% against 0.32%, so the frame representation is sufficient for backbone generation. Sequence-specific binding stayed at the rate of the docking pipeline in both, 0.031% and 0.027% against 0.027%. However, representing the contacting atoms explicitly, as the all-atom model does for the guidepost atoms and the atomized DNA residues they contact, left the rate of sequence-specific binding unchanged.

The atomic detail in that model is limited in two ways. It is confined to the seeded contacts as the seeds are supplied before generation from a rotamer library, the rest of the interface is represented by backbone frames, and the model must additionally assign

each guidepost atom to a residue position. The seeds also fix the recognition mechanism in advance, since the sampled contacts are arginine, asparagine and glutamine to guanine and adenine, so a design can read the bases only in the ways the seed library supplies.

This motivated RFdiffusion3, which represents every atom explicitly and denoises all-atom coordinates directly, following the AlphaFold3 architecture [1, 18]. Every atom is a token, so the model places the atoms that make an interface without a separate seeding step and without assigning guidepost atoms to sequence positions, and a design condition can be placed on any atom. That model, and the DNA-binder design and prediction pipeline built on it, are the subject of Chapter 3.

3 De Novo Generation of Sequence-Selective DNA Binders

Chapter 2 traced the DNA binder design problem through two generations of frame-based diffusion models, each of which improved the rate at which designs bound DNA while leaving sequence selectivity at the level of the previous pipeline. This chapter describes the design campaign that followed, built on RFdiffusion3, an all-atom diffusion model, and on an explicit *in silico* negative-design step using AlphaFold3. The campaign yielded specific binders for 7 of 15 DNA targets from 96 wet-lab tested designs per target, a rate roughly two orders of magnitude above the previous *de novo* efforts in our lab [37, 87]. I contributed to the development and *in silico* evaluation of RFdiffusion3 for DNA binder generation, to the construction of the design pipeline, and to the analysis of the resulting designs.

3.1 The Generative Model: RFdiffusion3

The models in Chapter 2 place a residue frame at each position and recover side-chain atoms from that frame through torsion angles or through atomization of selected residues. However, protein-DNA recognition depends on the positions of side-chain tip atoms relative to base edge atoms, so the atoms that determine specificity are the ones represented least directly. Therefore, RFdiffusion3 (RFD3) diffuses those atoms directly [18].

All-Atom Diffusion

The fundamental unit of the diffusion process in RFD3 is the atom. Because the amino acid sequence is unknown during generation, every residue is represented by a fixed 14 atoms: 4 backbone atoms and 10 side-chain atoms, 10 being the side-chain atom count of tryptophan, the largest canonical side chain. For residues with fewer side-chain atoms, the unused atoms are placed at the C_β position, or at C_α for glycine. Additionally, the terminal oxygen of serine and the sulfur of cysteine are assigned to different virtual atoms, which lets the network distinguish two residues that are otherwise identical in this representation. The model is built on the AtomWorks framework.

Generation therefore produces backbone and side-chain coordinates jointly. The model commits to a side-chain placement at every position while the backbone is still forming, so the backbone it converges on is one that can support those side chains. Therefore, the hydrogen bonding geometry at the protein-DNA interface can be specified as a design condition on individual atoms.

Architecture

RFD3 uses a transformer-based U-Net over atoms, following the AlphaFold3 diffusion module [1], in three parts (Figure 16): a downsampling module that encodes atomic and residue-level features together with the partially noised structure, a sparse transformer that processes token-wise information, and an upsampling module that modulates the atomic features with the token-wise features and predicts coordinate updates.

In order to sample backbone and side-chain coordinates jointly, the atom-level and

residue-level features must exchange information, and RFD3 introduces two mechanisms for this. Information transfer between atomic and token features runs through a series of sparse attention blocks in which residues and atoms attend to each other only when they are geometrically close in the current noised structure. This restriction focuses the network on local interactions and reduces overfitting. Pooling between the fine-grained atom representation and the coarser token representation is performed by cross attention, where each atom’s features are weighted by an attention operation before being pooled into the token features, since the token-level transformer can then attend to specific atoms within a token. The token features are read back out by cross attention with token splitting.

The input to a design model is a set of design conditions rather than an amino acid sequence, so the module that processes conditioning information can be much smaller than the corresponding module in a structure prediction model. The Pairformer is reduced from 48 layers to 2, and the triangle multiplicative update and triangle attention operations used in RoseTTAFold and AlphaFold are omitted entirely. The final network has 168M trainable parameters, against roughly 350M for AlphaFold3, and most of the compute is spent in the diffusion module. Therefore, RFD3 runs about an order of magnitude faster than previous versions of RFDiffusion over typical protein length ranges.

Conditioning

For binder design and motif scaffolding, the coordinates of the target or the motif are supplied as exact values inside the developing noise cloud, and the network refines

the diffused protein in that context. RFD3 was additionally trained to follow several conditions that an all-atom representation makes directly expressible (Figure 17):

- Hydrogen bond conditioning. Individual target atoms are labeled as hydrogen bond donors or acceptors, and the model is trained to make the corresponding interaction. For small molecule binder design this raises the fraction of designs making such interactions from 26.67% to 32.67%; for hydrogen bonds to DNA bases, from 11% to 11.3%.
- Solvent accessibility. Per-atom relative solvent accessible surface area labels specify the extent of burial of individual target atoms.
- Center of mass. The position of the generated protein’s center of mass relative to the target is specified by the placement of the initial noise cloud, through what RFD3 calls an ori token.
- Atom-level hotspots. Binding epitopes are specified at the level of individual atoms rather than whole residues.
- Symmetry. Symmetric noise at initialization produces symmetric structures.
- Unindexed atomic motifs. An additional token is appended to the diffused tokens for each motif residue, containing only the atoms being held fixed, and the model is trained to overlap one of the diffused side chains with those atoms. The sequence position that hosts the motif is chosen by the network during generation.

Adherence to these conditions is improved with classifier-free guidance [43]. At each

denoising step the network runs two forward passes, one conditional and one unconditional, and takes a weighted average of the two. This guidance raises the hydrogen bond fractions above to 36.67% for small molecules and 12.5% for DNA bases.

In RFdiffusion All-Atom the specified tip atoms are represented separately from the residue frames that carry the rest of the structure, and conditioning on hydrogen bond donor or acceptor status or on atomic burial is available only for that small subset of atoms represented explicitly [60]. However, in RFD3 a tip atom is a token like any other, distinguished only by having fixed coordinates, and a condition can be placed on any atom in the system.

Training

RFD3 was trained on the protein-protein, protein-small molecule, and protein-DNA complexes of the Protein Data Bank, on motif scaffolding tasks, and on a set of high-quality AlphaFold2 distillation structures. Each training structure is annotated with the features a designer might specify at inference: hydrogen bonds, binding partners, functional motifs. A training example is generated by noising a native structure to some extent and predicting the positions of the backbone and side-chain atoms of each residue, given some subset of the applicable conditioning information. Sampling that subset teaches the model to work from partial specifications.

Training proceeds hierarchically to limit overfitting across the range of tasks. The network is pre-trained on a mixture of AlphaFold2 predictions and PDB structures, then fine-tuned on a mixture weighted more heavily toward DNA complexes and protein-protein interfaces. PDB entries deposited through December 2024 were used. The

weights used for the design campaign in this chapter come from the same training trajectory as the released RFD3 checkpoint, trained for longer.

Inferring the Bound DNA Shape

In order to design against a DNA target, we require the bound conformation of that target, which is a property of the complex and therefore unavailable before the protein exists. RFD3 addresses this by predicting the protein structure and the DNA conformation jointly from the DNA sequence alone (Figure 17b).

The authors benchmarked this on three DNA sequences held out of the training set: 100 structures per sequence, four LigandMPNN sequences per backbone [26], then AlphaFold3 prediction. Design accuracy is measured by DNA-aligned RMSD, where the DNA phosphate atoms of the design and the prediction are aligned, and the RMSD is computed over protein C_α atoms with terminal loops trimmed. Under a 5 Å threshold, RFD3 passes 8.67% of monomeric designs and 6.67% of dimeric designs. Alternatively, if the RFD3-generated interface is held fixed during LigandMPNN sequence design, these fall to 6.5% and 5.5%. However, letting the model sample the DNA conformation reduces monomeric refolding success slightly relative to supplying a fixed DNA structure, although it samples a wider range of DNA conformations. For dimeric generation it improves refolding success in some cases.

Therefore, we held the target DNA fixed during diffusion for the campaign described below. The DNA structures were supplied by AlphaFold3.

3.2 A Pipeline for Sequence-Selective DNA Binder Design

The pipeline runs in two segments, which we call the binder block and the specificity block (Figure 18, Figure 19). The binder block generates backbones against a target DNA structure, fits sequences to them, and filters on agreement between the design model and the AlphaFold3 prediction of the designed complex. The specificity block takes the surviving designs, folds each one against every target in the panel, and ranks by the margin between on-target and off-target prediction confidence. The first segment selects for binding, while the second performs negative design against the other targets in the set.

Target Selection and Preparation

We selected six core targets to benchmark the pipeline, such that each unique 4-mer, and its reverse complement, occurs in at most one of the six (Table 1). Five are therapeutically relevant sites outside transcription factor binding space: two Oct4 guide RNA sites, the CAG trinucleotide repeat expanded in Huntington’s disease [6], the site of an epigenetic editor for LDL cholesterol reduction [92], and the site of a compact epigenetic editor silencing prion protein [77]. The sixth is the TATA box, which the native TATA-binding protein recognizes through the minor groove [56].

A further ten targets were selected without the k-mer constraint (Table 1). Seven are native transcription factor binding sites, taken from high-information content position weight matrices in JASPAR [20]. The other three are the human telomeric repeat [75] and two guide RNA sites against pathogenic DUX4 expression in facioscapulohumeral dystrophy [41]. These ten share sequence with one another and with the core set, which

makes the selectivity problem harder and closer to what a genomic target presents. For each JASPAR entry we took the monomeric information-rich core and padded it with a random combination of G and C to a total length of 10 or 12 base pairs, whichever was closer.

Each target sequence was folded into a double helix with AlphaFold3 (seed 42, single diffusion sample), and that structure was the starting point for design.

Binder Block

We generated backbones with RFD3 conditioned on the target DNA structure, which was held fixed throughout diffusion. Center-of-mass conditioning was applied through multiple ori tokens, one per six consecutive base pairs. For a given six-base-pair stretch of the double helix, the ori token is placed 3 Å from the centroid of that stretch, along the major groove direction and perpendicular to the helical axis. By sampling a range of ori placements, we produce a range of interaction modes across the target. For each ori, 5100 backbones of length 120 to 150 residues were generated, with hydrogen bond conditioning applied to candidate major groove donor and acceptor atoms [23] and the `is_non_loopy` feature set to true.

We then fit sequences to the backbones. Each diffused backbone was relaxed with Rosetta FastRelax [3], and LigandMPNN sampled 5 sequences per backbone at a sampling temperature of 0.1 [26]. We use LigandMPNN sequences rather than the sequences RFD3 itself produces because AlphaFold3 predictions of the LigandMPNN sequences land closer to the designed complex structure.

Each sequence was folded with AlphaFold3 in complex with its on-target DNA, without templates. Designs whose predictions fell within 8 Å DNA-aligned protein C $_{\alpha}$ -RMSD of the design model were carried into a resampling round, where LigandMPNN sampled 5 further sequences per design at temperature 0.1, and those were folded again with AlphaFold3. This sequence design and structure prediction cycle corresponds to one step of the optimization used in ProteinHunter and HalluDesign [24, 31], and it improved the *in silico* metrics substantially.

Designs leaving the binder block were required to fall within 3 Å DNA-aligned protein C $_{\alpha}$ -RMSD of the design model, to have interface predicted TM score (ipTM) above 0.7, and to make a large number of protein-DNA hydrogen bonds. Interaction counts were computed with DSSR v1.7.8 [67]. Total hydrogen bonds are the count of all bonds DSSR annotates as **nt:aa**; major groove hydrogen bonds are the subset made to base atoms exposed on the major groove edge; supporting hydrogen bonds are protein-protein (**aa:aa**) hydrogen bonds, excluding those involving the protein backbone, made to residues that themselves hydrogen bond the DNA.

Supporting hydrogen bonds select for preorganized side-chain conformations at the interface. A side chain held in place by an intraprotein buttressing network pays less conformational entropy on binding its intended base, and the same restraint narrows the set of alternative rotamers available to it for reading a different base, so buttressing contributes to selectivity as well as to affinity.

Specificity Block

Designs entering the specificity block were required to have minPAE below 1.25 at the end of the binder block. Each was resampled with LigandMPNN at 100 sequences per backbone, temperature 0.1, and the resulting sequences were folded with AlphaFold3 against their on-target DNA. Passing designs fell within 1.5 Å DNA-aligned protein C_α-RMSD and had ipTM above 0.9.

Surviving designs were then folded all-by-all, in complex with its on-target sequence and with the six core targets, plus the ten additional targets when the design’s on-target belonged to the additional set. These predictions were templated, using the most recent AlphaFold3 prediction of that design as the template for the protein chain, which fixes the protein conformation across the panel so that the confidence differences reflect the DNA.

The selection metric is built from the minimum predicted aligned error over protein-DNA residue pairs. We write P for the set of protein residues and D for the set of DNA residues,

$$\text{minPAE} = \min_{i \in P, j \in D} \text{PAE}(i, j),$$

and the on-target margin is the difference between the best off-target minPAE and the on-target minPAE (Figure 20):

$$\Delta \text{minPAE} = \min_i (\text{minPAE}_{\text{off}_i}) - \text{minPAE}_{\text{on}}.$$

Designs were ranked in decreasing order of ΔminPAE and the top 96 per target were ordered as synthetic genes.

minPAE was established as an *in silico* filter for designed binders against protein targets [110], and two benchmarks on native proteins indicate that it is also informative for DNA targets. On a subset of transcription factors from UniProbe [45], predicted in complex with their recorded consensus sequences and with mutated derivatives, minPAE separated binders (PWM score above 0) from non-binders with a mean per-transcription-factor AUC of 0.69, whether or not multiple sequence alignment information was supplied. On a mutational library of the *Saccharomyces cerevisiae* bHLH transcription factor Pho4 binding E-box variants, with binding defined as apparent K_D below 10 μM [2], minPAE reached a mean per-sequence AUC of 0.75 in single-sequence mode and 0.51 with an MSA supplied. Pairs correctly flagged as likely non-binders by high minPAE in single-sequence mode shift to low minPAE once the MSA is added, producing false positives. The evolutionary constraints supplied by the MSA appear to reduce the contribution of the subtle geometric and chemical interface differences that distinguish point mutants. Single-sequence performance is the relevant regime here, since *de novo* designs are folded without an MSA.

3.3 In-Silico Results

All-Atom Generation Outperforms Residue-Level Generation

We compared four generation methods on the same downstream pipeline: CA-RFdiffusion, RFdiffusion All-Atom, BoltzGen, and RFD3 [18, 60, 63, 90]. CA-RFdiffusion and RFdiffusion All-Atom were the fine-tuned DNA-binder variants described in Chapter 2.

The former fine-tunes RoseTTAFoldNA for the diffusion generative task on protein-DNA interfaces cropped around a protein contact, dropping the loss on backbone frame rotations to yield a C_α -only model with a subsequent refinement step [13]. The latter fine-tunes RoseTTAFold All-Atom on tip-atom placement and motif scaffolding, seeded by RIFGen-sampled major groove rotamers [19].

For each method we ran 1000 diffusion runs against each of the six core targets, generated 5 LigandMPNN sequences per backbone, and predicted each with AlphaFold3. The atomic interaction filters and the LigandMPNN-AlphaFold3 cycling were omitted, since the comparison was intended to isolate the generator.

RFD3 produced the highest fraction of designs whose AlphaFold3 predictions agreed with the design model by DNA-aligned RMSD (Figure 18c). BoltzGen, which is likewise an all-atom diffusion model, ranked second and was well ahead of both the residue-level CA-RFdiffusion and the hybrid residue-atomic RFdiffusion All-Atom. RFD3 and BoltzGen sampled comparable backbone diversity among passing designs, with BoltzGen covering more distinct clusters at the most stringent refolding cutoff and RFD3 covering more overall.

The interaction analysis indicates the source of the difference (Figure 23). The number of hydrogen bonds between protein side chains and DNA — to base atoms and to the phosphate backbone alike — was considerably higher for the two all-atom methods than for the residue-level ones, and approached the interaction density measured in crystal structures of native transcription factors bound to their sites (357 JASPAR structures with mean information content above 1.5). Chapter 2 attributed the specificity failures

of the frame-based campaigns to sparse hydrogen bonding at the interface, and the all-atom models raise the count of these bonds to native levels.

Δ_{minPAE} Predicts Selectivity

The binder block selects only for on-target binding. A design that wraps a target over ten base pairs and makes native-like numbers of base contacts may make comparable contacts to a different sequence, and the binder block scores both the same.

Across the binder block hits ($n = 59$), Δ_{minPAE} correlates with the experimental selectivity margin, defined as the difference between the on-target and strongest off-target normalized binding signal (Figure 22). Among the designs satisfying $\Delta_{\text{minPAE}} > 0$, the criterion actually applied in the specificity block, 23 of 32 (71.8%) proved on-target selective experimentally.

3.4 Experimental Characterization

We tested designs from both blocks. For each of the six core targets we characterized the top 96 specificity block designs by Δ_{minPAE} , assembling genes individually from oligo pools, cloning them into a yeast display vector, and transforming into *S. cerevisiae* EBY100. To measure what the specificity block contributes, and to compare against approaches lacking *in silico* specificity filtering, we characterized roughly 1,200 designs per target from the binder block alone, synthesized as an oligonucleotide pool and screened by fluorescence-activated cell sorting followed by deep sequencing. Designs from the specificity block carry the prefix DBS and those from the binder block, DBB.

Binding and Selectivity by Yeast Display

Binding was quantified by flow cytometry as the normalized PE/FITC signal, the DNA binding fluorescence relative to protein surface expression, in the presence of 1 μ M fluorescent on-target or off-target DNA for each protein-DNA pair. A design counts as specific when the normalized signal is higher for its on-target than for the five off-targets. Binder block designs identified as enriched by sorting and sequencing were subsequently characterized as individual clones in the same flow cytometry assay, which allows the two blocks to be compared on the same measurement.

From the binder block pools we identified 42 specific binders across 5 of the 6 core targets, an average success rate of 0.46% over all binders tested. From the specificity block designs we identified an on-target binder for 4 of 6 core targets and a specific binder for 3 of 6, with per-target success rates from 1/96 (Oct4-gRNA2) to 13/96 (PRNP-site) and an average near 3%. The *in silico* specificity filter therefore raised the frequency of specific designs about 6-fold, from 0.5% to 3%, and the combined protocol produced specific designs at a rate roughly 100 times higher than our previous *de novo* DNA binder design effort [37]. Among the specificity block designs, most of the designs that bound at all were also specific. In total we obtained 59 specific binders across 5 of the 6 core targets, 42 from the binder block and 17 from the specificity block (Figure 21).

We then tested the pipeline against the ten additional targets. Designs for each were filtered *in silico* against all 15 off-targets, except HSTelo, which was filtered against the six core targets only. We obtained synthetic genes encoding 96 designs for 9 of these 10 targets, selected in decreasing order of Δ minPAE and sampling up to 5 designs per

diffusion backbone. The Ebox target was skipped due to low *in silico* pass rates, possibly a consequence of its palindromic symmetry. Flow cytometry identified specific binders for 4 of the 9 targets: HSTelo, the binding site of the homeodomain transcription factor Engrailed, the NF κ B site, and the Dux4 gRNA site. Binding rates ranged from about 1% to about 19% and specificity rates from about 1% to about 14%, and as with the core set, most designs that bound their target were also specific.

Specificity filtering thus produced 31 specific binders across 7 targets from 96 designs tested per target, and combining both blocks gives specific binders for 9 targets in total. These 9 span nearly as wide a range of DNA sequence diversity as the targets of native transcription factors, measured as the spread of Evo-1 embeddings of the target sequences projected into a principal component space fit on JASPAR vertebrate transcription factor sites [78].

However, designs filtered against all 15 off-targets expressed at lower rates on the yeast surface than designs filtered against 5 or 6, which indicates a tradeoff between folding and selectivity that will need to be handled directly as the off-target panel grows.

To calibrate these selectivity levels against native proteins under the same conditions, we assayed 39 native DNA binding proteins recognizing sites in our target set. Two expressed well on the yeast surface and bound their target bound their target, the Engrailed homeodomain and the TATA-binding protein. The homeodomain was specific, with slight off-target activity toward the NF κ B site, matching the behavior of our design DBS29 against the same site. TBP bound several off-target sites strongly, likely because the short DNA targets in this assay are malleable to the deformation TBP induces on

binding (PDB 1TGH). The specificity of our designs falls within the range these native proteins show under the conditions of the assay.

Affinity and Biophysics

For biophysical characterization, 82 designs were expressed in *E. coli* and purified by immobilized metal ion affinity chromatography followed by size-exclusion chromatography. Of these, 46 expressed solubly and were monomeric by SEC. Biolayer interferometry gave equilibrium dissociation constants for the cognate site ranging from 3 to 30 nM (Figure 24), with kinetic parameters determined by global fitting to a 1:1 binding model across the dilution series.

Additionally, the specific designs made dense protein-DNA interatomic interactions, consistent with the levels seen in native transcription factors and exceeding those of DNA binders from our previous design pipelines (Figure 23).

Single-Base Specificity and Recognition Mechanism

Fine-grained specificity was measured with a yeast display competition assay [37]. Cells were incubated with 1 μ M biotinylated on-target DNA in the presence of 8 μ M unlabeled competitor DNA carrying a single base substitution, and the binding signal for each variant was measured relative to the original target. We read the results against the design models to identify the interactions responsible for each position's contribution (Figure 25, Figure 26).

DBB5, against the almost fully poly-purine PRNP-site, has two helical subdomains contacting different regions of the target, connected by an extended region outside

regular secondary structure. Asn114 makes a bidentate interaction selecting A6. The design binds with a K_D of 3 nM. DBB3, against the same site, binds at 10 nM and its HT-SELEX profile shows specificity toward the 5-mer GAGGA, which occurs twice in the target (Figure 27). DBS5, a specificity block design against the same site, prefers polypurines with guanine preferred over adenine, consistent with the abundance of arginine side chains at its interface, and over positions 1–10 it prefers the intended sequence over 35 of the 40 single base variants tested.

DBS14, against the Oct4-gRNA2 site, illustrates the buttressing that the supporting-hydrogen-bond filter selects for. Arg58 makes a bidentate interaction with G3 and is buttressed by a hydrogen bond from the adjacent Tyr123. Additionally, Tyr123 makes a supporting polar contact with Gln127, which reads T5. Arg82, Arg7, and Arg5 hydrogen bond multiple bases in the minor groove, which contributes to affinity and to specificity [84, 102]. The design binds at 6 nM and discriminates strongly at positions 1 through 7, where only 2 of 28 mutations give stronger signal.

DBB26 targets the AT-rich TATA box through base-specific polar contacts in the major groove. A8 and A4 are read by the bidentate amine side chains Asn50 and Asn122, each buttressed by further polar side chains, and Arg53 contacts a G12-T13 dinucleotide step. The native TATA-binding protein reads this same site primarily through the minor groove, so DBB26 establishes that the sequence is also addressable from the major groove, at 30 nM affinity and with on-target selectivity.

DBS15, against the CAG repeat, adopts an integrated extended single-domain topology with extensive hydrogen bond networks. Asn37 hydrogen bonds a G15-C16 step within

a buttressing network. Arg47 and Arg62 make bidentate interactions with G18 and G21. Multiple hydrogen bonds to DNA phosphates stabilize the bound DNA backbone conformation, supporting indirect readout of the target [89]. Affinity is 10 nM. DBB32, against the same target, recognizes the repeating unit GCA, which overlaps the reverse-complement repeat TGC offset by one base on the opposite strand, a pattern consistent with dimeric or multimeric binding (Figure 27).

DBS18, against the human telomeric repeat, contacts a G4-T5 step with Arg51 and an A7-G8 step with Gln120 and Arg124, and extends a loop carrying several positively charged side chains into the minor groove. It binds at 4 nM. The competition assay shows strong discrimination between positions 2 and 8 and weak discrimination over the last four positions, where the extended flexible loop contacts the DNA backbone alone and where the AlphaFold3 pLDDT of the contacting residues is low. DBS20, against the same target, engages the site through its major groove interface alone, binds slightly more weakly, and discriminates single mutations better, as it achieves very high specificity for the central GTTA, consistent with a dense network of side-chain-base interactions engaging both strands there, and sustains more modest selectivity across the full 12 base pair site, with only 4 of 36 single-site variants binding more tightly than the original.

DBS22, against the NF κ B site, is monomeric where the native factor is a dimer (PDB 1NFK), and uses two helical subdomains to make extensive base contacts. Arg43 contacts a G15-G16 step; Arg95 and Thr92 jointly read T7 and G17; Gln90 makes a bidentate interaction with A5 within a network of polar side chains and DNA backbone

contacts; Arg95 and Ser91 contact A18 and T6. The competition assay shows strong specificity for the central ATTC.

DBS29, against the homeodomain site, has one subdomain that reads as an extension of the homeodomain fold with three supporting helices in place of two. Asn140 and Gln143 make bidentate interactions with A6 and A18, and the design makes extensive contacts to the narrow minor groove of the AT-rich core [83]. Its second interface hydrogen bonds five consecutive phosphate moieties. DBS29 binds at 13 nM and is highly specific despite having low average AlphaFold3 pLDDT and only two positions read by major groove bidentate interactions, which points to indirect readout of the distinct shape and flexibility of the AT-rich target through backbone hydrogen bonds and minor groove contacts.

Across the designs characterized this way, the majority have at least seven positions showing base-specific behavior under single mutation. The variation in specificity between positions within a design tracks two quantities. Where it reflects contact coverage, the fixed-dock specificity predictors NA-MPNN and DeepPBS, which take the folded structure as input, broadly reproduce the observed single-base effects for 4 of the 5 designs examined, including some off-target preferences such as DBS20’s preference for C at position 9 [61, 74]. Where it reflects local rigidity, the average AlphaFold3 pLDDT of the contacting residues at each base pair position is the better indicator, since pLDDT is lower over loopy regions and regions with few contacts, and those are the positions where selectivity is weakest, including DBS14 positions 8–10, DBS29 positions 1–4, DBS5 positions 4 and 12, and DBS18 positions 9–12. However, AlphaFold3 minPAE

was uninformative about variant selectivity, which bounds the resolution at which the metric used to select these designs operates.

HT-SELEX against a randomized library gives a more global picture of specificity than a single-mutant panel, and it is the standard measurement for native transcription factor specificity [51]. We ran it on several binder block designs. DBB5 shows a strongly informative 8-mer motif with 5 positions aligning to the intended PRNP-site; DBB26, against the AT-rich TBP-site, gives a 10-mer with 7 of 10 positions matching; DBB20, against Oct4-gRNA2, gives a 10-mer with 6 of 10 closely matching; DBB27, against Oct4-gRNA1, gives a shorter 6-mer with 4 positions matching (Figure 27); DBB31, against the CAG repeat, gives a longer profile with more modest information content, matching the intended target as top or second-best preference at seven positions. However, these designs were screened *in silico* against a panel of six to fifteen sequences rather than for broad specificity, and they nonetheless return information-rich binding motifs.

Novelty

The designed sequences fall outside the sequence space of known proteins. Projected into a principal component space of ESM-2 embeddings fit on 532 protein chains taken from crystal structures of native protein-DNA complexes, the 73 designed sequences occupy a distinct cluster with broad spread inside it [64].

Additionally, the structures are globally distinct from native DNA binding proteins. TM-align against the PDB returns best matches with TM-scores around 0.5 to 0.6, at or near the 0.5 threshold conventionally taken to indicate fold-level similarity [111]. At the same time the designs share local features with native transcription factors —

extended protein-DNA interfaces and clusters of base-contacting hydrogen bonds of the kind seen in homing endonucleases [11, 22] — and their affinities are comparable to native levels. Some designs adopt modular helical bundle architectures resembling homeodomain proteins joined by flexible linkers, and others form integrated monolithic folds distinct from those of most natural transcription factors.

Extensions

Two modifications extend the approach beyond the standard case.

The first targets DNA in a non-standard conformation. To design minor groove binders against a bent TBP-site, we obtained the target structure by folding CGTATAAACG with the native transcription factor in AlphaFold3, then modified the binder block, using diffused protein lengths of 150–220, one ori per 8-mer placed perpendicular to the local helical axis and 1.5 Å into the minor groove direction, and hydrogen bond annotations on minor groove base atoms. LigandMPNN sampled 10 sequences per backbone in the first round and 50 in the second. The AlphaFold3 pass rate was 0.67%, well below the major groove case. Final cutoffs were 3 Å DNA-aligned protein RMSD, ipTM 0.7, 14 total hydrogen bonds, 10 supporting hydrogen bonds counting protein backbone bonds, minPAE 1.25, and at least 2 minor groove hydrogen bonds. Of 46 designs passing these criteria and tested experimentally, 2 bound (Figure 28).

The second extends the interface. Homing endonucleases achieve high specificity by reading about 20 base pairs, against the 7 to 9 base pairs read here, so extending the interface is a direct route to higher information content. We ran a small design cycle for the p53-site and NF κ B-site targets across a range of length specifications. Pass rates by

DNA-aligned protein RMSD fall with length but remain at 5–6% for lengths of 240–270. minPAE pass rates also fall with length, though at the most stringent cutoff (below 1.0) the 150–180 range performs best. The average AlphaFold3 pLDDT of the protein in complex with DNA rises with length, while the pLDDT of the same designs predicted as protein alone stays constant across length ranges, which indicates that the gain is in the complex.

3.5 Future Work: Latent Diffusion and Inference-Time Search

The generative models used so far in this thesis diffuse coordinates directly. CARFdiffusion diffuses alpha-carbon positions, RFdiffusion All-Atom diffuses residue frames together with individually represented atom nodes, and RFD3 diffuses the 14 atoms of every residue. La-Proteina represents the atomic detail in a learned latent instead [36]. Figure 1 of that work depicts the encoder, the decoder, and the joint flow model, and the reader is referred to it for the details of the framework. A variational autoencoder encodes each residue’s amino acid identity and its non-alpha-carbon atom coordinates into a per-residue latent variable of fixed dimension, and a flow matching model generates the alpha-carbon coordinates and those latents jointly. A frozen decoder then recovers the full-atom structure and the sequence from the generated alpha carbons and latents. Since the latent has the same dimension at every residue, the size of the representation is independent of amino acid identity, and the denoiser is a plain transformer, omitting the triangle multiplicative and triangle attention operations. Generation is therefore fast enough to supply the thousands of candidates per target that a design campaign requires.

Proteina-Complexa extends this framework to binder design [28]. In Proteina-Complexa, the flow model is conditioned on the target, and the autoencoder encodes and decodes monomeric binders, so one autoencoder serves every target type. Additionally, the target is featurized per residue, embedded, and concatenated to the binder tokens, and the pair representation is formed over binder and target together. The model co-generates the binder sequence with its structure, so the decoder outputs the sequence directly. We have trained autoencoders on DNA in addition to protein and trained flow matching design models with them. This puts the target in the same latent representation as the binder, and the flow model can therefore reason over protein and DNA latents jointly while designing.

In the pipeline of §3.2, each design is sampled independently and filters are applied at the end to finished designs. This is because the model, while being an all-atom model, does not generate high-quality codesigned sequences — instead, ProteinMPNN must be run on the resulting backbones to generate high quality sequences. This is because the all-atom diffusion objective forces the model to decide early on in the diffusion process what the final sequence of the protein it is generating will be, making it difficult for the model to modify the sequence mid-stream as this means several of the atoms in the atom14 representation must be moved jointly to form a new amino acid residue. However, with Proteina-Complexa, because the sequence and sidechain information is compressed into a fixed-dimension latent, it is much easier for the model to refine the sequence of the structure it is generating during the course of the flow-matching process. This yields better codesign sequences, obviates the need for a sequence-design step, and enables powerful inference-time search and guidance techniques to explicitly guide

generation toward optimizing metrics of interest. In search, a score can be computed on a partially generated design and used to decide which trajectories continue. Figure 1 of Didi et al. depicts both the target-conditioned generation and the beam search over the generative trajectory, and the reader is referred to it for the details of the search procedure [28]. Complexa maintains a set of partially denoised trajectories, branches each into several new stochastic trajectories, runs them forward a fixed number of steps, decodes each candidate to an atomistic binder, co-folds it with the target, and keeps the highest-scoring candidates as the next beam. Additionally, explicit inference-time guidance is possible, using a gradient over the structure and the latents coming from a classifier that is trained to predict final design quality from a partially generated design.

As a proof of principle, we have run search against the six core targets of §3.2, with the on-target interface minPAE as the reward (Figure 29). However, the effect depends on the target. On the PRNP-site the fraction of designs whose AlphaFold3 prediction agrees with the design model rises from 4.7% to 40.1%, and on the TBP-site from 11.5% to 20.9%, while on Oct4-gRNA2 it falls from 14.6% to 0.6%. The search may be committing early to whichever region of the sample space the initial beam happens to cover, although we have not tested this in detail.

The score used for search may also carry an off-target term. This moves the negative design of the specificity block from a ranking applied to finished designs into the generation itself. We attempt a proof of concept where we penalize a low minPAE against a competing target and gate that penalty on the on-target minPAE first crossing a threshold, since the penalty is only meaningful once a trajectory has found something

that binds the intended target. Furthermore, the whole reward is switched on halfway along the trajectory, as every evaluation costs an AlphaFold3 co-fold. Figure 30 shows the outcome for the target pairs we have run so far. Against the PRNP-site with the TBP-site as the off-target, 48% of the searched designs are specific *in silico*, above the 34% of BoltzGen and the 19% of RFD3, while against the TBP-site with the PRNP-site as the off-target every searched design binds both. Taken together, these search experiments demonstrate the potential of a latent diffusion model to further improve the efficiency of the design pipeline for DNA binding proteins. Further work is necessary to improve the underlying flow model, and to explore other search and guidance strategies.

3.6 Discussion

Sequence-specific DNA recognition is a harder design target than protein-protein binding, because DNA sequences share a common global structure where protein targets present diverse shapes. On-target positive design, which usually suffices to obtain specific protein binders without an explicit negative design step, produced designs here that bound many sequences, since optimizing for binding affinity alone selects for the extensive backbone and shape complementarity that any B-form DNA offers.

Three elements together produced selectivity in this campaign. Diffusion generated proteins that engage the target over an extended region, which makes many base-specific contacts available. Explicit negative design removed candidates predicted to bind the off-target panel, widening the energy gap between the intended complex and its competitors [33]. Additionally, filtering on hydrogen bond networks selected for direct readout through contacts to the bases and for indirect readout through contacts to the

phosphate backbone, which reads the shape differences between sequences [83].

Four limits remain. Selective binders were obtained for 9 of the 16 targets attempted, and 7 returned none. Of those 7, five were tested with 96 designs each, one was a core target tested with 96 specificity block designs and roughly 1,200 binder block designs, and the Ebox target was skipped, as too few of its designs passed the *in silico* filters. Specificity here is measured relative to the 6- or 15-member off-target panel used in filtering, which covers a small fraction of the sequence space a genome presents. Expression rates fall as the off-target panel grows, so the folding-versus-selectivity tradeoff tightens on the way to genome-wide specificity, in the manner expected of any protein optimized against competing objectives [16]. The selection metric also operates at the resolution of whole targets, as ΔminPAE enriches for designs that discriminate among a panel of distinct sequences, while AlphaFold3 minPAE is uninformative about the effect of a single base substitution. Additionally, within the ordered set, the designs that bound and the designs that bound specifically have the same minPAE distribution as the set as a whole below 1.15 (Figure 31), so ranking harder on minPAE past that threshold does not enrich for binding or selectivity.

Several routes address these limits. We could optimize directly for specificity during generation, through the inference-time search of §3.5, which would allow the tradeoff between folding, binding, and selectivity to be balanced explicitly rather than imposed as a filter after the fact. Additionally, we could use orthogonal structure prediction methods during iterative design and final evaluation, as in ProteinHunter and HalluDesign, which would reduce overfitting to a single predictor [24, 31]. We could also extend the interface,

either by lengthening the diffused chain or by combining designs through rigid fusions or engineered heterodimeric interfaces, which would raise the information content of the recognized site toward the ~ 20 base pairs at which homing endonucleases operate. Finally, characterizing target specificity in cells by ChIP-seq is the necessary next measurement, since the yeast display panel establishes relative selectivity among a handful of sequences rather than occupancy across a genome.

I take the prediction gap to be the limit that most constrains the next stage of this work. Every filter in this pipeline is an AlphaFold3 confidence metric applied to a structure AlphaFold3 itself predicted, and the analysis above bounds what that provides. MinPAE discriminates cognate from non-cognate targets across a panel with a per-transcription-factor AUC near 0.7, and is uninformative at single-base resolution. Two methods that do reproduce single-base behavior, NA-MPNN and DeepPBS [61, 74], require a well-folded input structure and return a specificity profile where the pipeline needs a binding decision. The inference-time search of §3.5 also uses AlphaFold3 and is therefore bounded by the quality of AlphaFold3 as an oracle. Therefore, improving our ability to filter designs, especially among the many false positives that AlphaFold3 minPAE predicts to be binding but do not work experimentally, would improve our ability to generate DNA binding proteins for difficult targets with increased selectivity. This effort is the subject of Chapter 4.

4 Discriminative Models for DNA Binder Prediction

Chapters 2 and 3 describe the generative stage of the design pipeline, which proposes candidate proteins. This chapter describes the discriminative stage, which predicts whether a designed protein binds a given DNA sequence and ranks a panel of DNA sequences for one protein. Across four design campaigns the rate at which designs bound any DNA sequence rose by more than an order of magnitude, and sequence-selective binding improved in the final campaign, where we applied an explicit *in silico* selectivity filter for the first time. In each campaign the *in silico* filter correlated weakly with the wet-lab outcome it was chosen to predict. On a prospective orthogonality assay of 48 designs, each measured against a panel of 16 DNA targets, a model I built ranks the panel for a fixed design at a mean per-design Spearman correlation of 0.46, against 0.01 for the AlphaFold3 confidence metric the pipeline currently uses. On the five DNA targets withheld from training entirely, the model reaches 0.60 against 0.25 for AlphaFold3. Below, I describe the data curation, the model, and its evaluation.

I curated the training data, built and evaluated the model described here, and performed the retrospective analyses of the earlier campaigns. The experimental data used to train and evaluate the models comes from the design campaigns of Chapters 2 and 3.

4.1 Motivation

We have run four campaigns against DNA targets with the same overall pipeline: generate backbones, fit sequences to them, filter *in silico*, and screen the survivors in the wet lab. Each replaced the generative model, the filter, or both (Table 2).

Across the first three campaigns, changing the generative model raised the rate at which designs bound DNA by more than an order of magnitude, but the rate of sequence-selective binding was not improved. The fourth campaign added an explicit selectivity filter, ΔminPAE , which raised the selective rate to roughly 3% and gave specific binders for 7 of 15 targets from 96 designs tested per target (Chapter 3). Across both design blocks we obtained specific binders for 9 of the 16 targets attempted, at per-target specificity rates from about 1% to about 14%.

Currently, we have used three different *in silico* filters across these campaigns. Rosetta interface metrics, which count hydrogen bonds to bases and bidentate contacts, were uninformative about experimental selectivity. RoseTTAFoldNA metrics, used retrospectively, raised the fraction of binders from roughly 7% to roughly half on the RIFdock design set, but then were inversely correlated with the experimental outcome on diffusion-generated designs, where the selective binders scored worse than the nonbinders on all four metrics: higher interface predicted aligned error, higher ligand RMSD, lower predicted probability of binding, and lower pLDDT (§2.3). We therefore adopted AlphaFold3 on a second retrospective analysis, and then used it as an optimization objective, resampling designs by partial diffusion and redesigning them until they passed AF3 refolding RMSD cutoffs. Unfortunately, although every ordered sequence passed those cutoffs, the selective binding rate did not improve (§2.3). In the RFDiffusion3 campaign, we applied minPAE as our optimization metric, and used an explicit specificity-based selection criterion where we sought to maximize the difference in minPAE between the on-target and the best predicted off-target (§3.3). This led to successful design of proteins selective or specific for 7 of 15 attempted DNA targets.

However, two limitations remain. First, beyond a certain threshold ($\text{minPAE} < 1.15$), minPAE no longer discriminates between nonbinders and binders (§3.6, Figure 31). Second, for 7 of the 16 attempted DNA targets, no successful design could be found. Therefore, a more powerful discriminative model, that works for more DNA targets, appears necessary.

AlphaFold3 Lacks Good Negative Data

The protein-DNA portion of AlphaFold3’s training data is the DNA-binding proteins of the Protein Data Bank, roughly 12,000 complexes drawn from roughly 200 families or sequence clusters, together with a transcription factor distillation set of 1,000 proteins and 16,000 complexes. These complexes compose the positive binding example set. Negative, non-binding examples are constructed synthetically by pairing a random protein monomer with a random DNA sequence, with the ground truth structure placing the protein randomly far away from the DNA. Separating those pairs from deposited complexes requires merely recognizing whether a protein is a DNA-binding protein, which is a different task from recognizing whether it binds a particular site. The authors sought to improve site-discrimination by augmenting the training data with padded random flanking DNA on either side of the target, which forces the model to locate the correct binding site within a longer stretch of DNA. However, protein family identity, which can be inferred from the sequence and the MSA, correlates with binding motif across the roughly 200 clusters, so a model reaches high accuracy by recognizing the family and recalling that family’s preferred sequence. Furthermore, whatever signal remains in the confidence metrics appears to track the depth and quality of the multiple

sequence alignment, which is a property of the protein rather than of the interface.

Selectivity Determinants in a Predicted Structure

Protein-DNA selectivity is mediated by direct contacts to base edges in roughly 60 to 70% of characterized cases, by readout of sequence-dependent DNA shape in roughly 20 to 30%, and by ordered water molecules bridging protein and base in a smaller number (Chapter 1) [68, 83]. A model that takes a predicted protein-DNA interface as input would have direct access to the base-edge contacts. However, it would have only indirect access to shape, since the DNA conformation it receives is the conformation a structure predictor assigned to that sequence, which is a product of the predictor’s own sequence-to-shape model rather than a measurement. Ordered water, however, is absent from predicted structures altogether. The model described below therefore seeks to learn from base-contact chemistry to predict binding.

Requirements

The model must meet the following requirements.

- The model re-ranks the designs that survive AlphaFold3 folding and filtering, rather than replacing the folding step.
- It ranks DNA sequences for a fixed protein, which is selectivity, and proteins for a fixed DNA sequence, which is affinity.
- It is light enough to evaluate every pair in an all-by-all panel, which is quadratic in the number of designs and targets.

- It generalizes across protein families rather than recognizing them.
- It generalizes from native transcription factors, which are most of the available data, to *de novo* designs, which are the intended application.

4.2 Data Curation

Assays

We collect training data for protein-DNA binding specificity that originates from a variety of assays. Below, I describe the main assay types and their readouts.

SELEX-seq incubates a protein with a randomized DNA library, separates bound from unbound sequences, amplifies the bound pool, and repeats under increasing stringency [51]. The readout is enrichment across rounds, which orders sequences by affinity for one protein on a relative scale.

Protein binding microarrays hybridize a tagged transcription factor to a double-stranded DNA microarray and detect bound protein with a labelled antibody, measuring every probe on the array in a single experiment [15]. Probes that show no signal are measured non-binders, which makes UniProbe the strongest source of negatives among the public datasets [45].

ChIP-seq crosslinks proteins to genomic DNA in a living cell, immunoprecipitates with a protein-specific antibody, and sequences the recovered fragments [112]. It reports occupancy *in vivo*, which is a function of chromatin accessibility, cofactor availability and competition as well as of intrinsic affinity, and some peaks reflect indirect binding through protein-protein interactions.

MITOMI measures protein-DNA interaction on a microfluidic device, recovering kinetics and equilibrium affinities for a small number of proteins against many sequences [35]. It supplies measured affinities rather than derived scores.

Quantitative TALE profiling assayed 21 representative TALE proteins against 5,000 to 20,000 unique DNA sequences each [82], giving binary binder and non-binder labels for a protein class whose recognition is modular and well characterized.

The in-house orthogonality assays of Chapters 2 and 3 measure each design against every DNA target in a panel by yeast display, giving a normalized binding intensity per design-target pair. They are the only source containing *de novo* designed proteins.

Position Weight Matrices as Labels

Unfortunately, nearly every public source is distributed as a position weight matrix rather than as the underlying reads. A matrix is built by aligning the binding sites recovered from an assay, counting base frequencies at each position, and converting those counts to log-odds scores against a background, such that any candidate DNA sequence can then be scored against this scores matrix (Figure 32).

I therefore generated the training examples from the matrices. I scored candidate DNA sequences against each protein’s matrix, took high-scoring sequences as binders and low-scoring sequences as non-binders, and used the score itself as a graded label. Recovering the raw reads from the source experiments would give measured labels in place of inferred ones, but this would be expensive relative to the expected gain. However, a position specific scoring matrix is calibrated against one protein’s own background,

so its scores compare two DNA sequences for the same protein, and each protein has its own scale. These labels therefore support relative ranking of DNA sequences for a protein, which is the objective used in §4.3, rather than absolute scoring of binding affinity across proteins.

Data Sources and Composition

Table 3 lists the curated datasets, with their sizes and readouts.

`dna_pwm` is the consolidated PWM-derived data, assembled from five public and collaborator sources: CisBP, JASPAR, HT-SELEX data contributed by the Taipale lab, ChIP-Atlas and UniProbe [20, 45, 51, 101, 112] (Table 4). Additionally, I examined HOCOMOCO and excluded it, since its entries overlap the sources already included [94].

The PWM data sources were combined and then duplicate proteins were removed. Additionally, every protein was balanced to approximate an even split between binders and non-binders when its negatives were generated. The prevalence of binding in the training set is therefore a construction choice, and I return to its consequences in §4.5.

`inhouse` is the orthogonality panel from the earlier campaigns in Chapter 2, and `inhouse_rfd3` the panel from the RFdiffusion3 campaign of Chapter 3.

Splits

I drew the held-out sets from the two in-house datasets, since the intended application is *de novo* designs. My first split held out ten whole proteins from each panel and left the DNA panel shared between the training and validation sides. However, every design

in one of these panels is measured against the same small set of DNA targets, so a held-out protein is still scored against targets the model has seen thousands of times in training, and a model could reach a good per-protein ranking by recalling which targets are generally bound. I therefore replaced this split with a two-dimensional block.

The second split holds out ten proteins and five DNA sequences from each panel, and trains only on those pairs whose protein and whose DNA both fall on the training side. This leaves four quadrants, of which one is the training set and three are held out (Figure 33).

I refer to the doubly novel quadrant, in which a held-out protein is measured against a DNA sequence never seen in training, as the evaluation set, and I report the two off-diagonal quadrants separately, since they separate generalization along the protein axis from generalization along the DNA axis. For `inhouse_rfd3` the training quadrant contains 690 rows over 92 proteins and 11 DNA targets and the doubly novel quadrant 50 rows over 10 proteins and 5 DNA targets, and for `inhouse` the corresponding quadrants contain 516 rows over 63 proteins and 16 DNA targets, and 45 rows over 10 proteins and 5 DNA targets.

I selected the five held-out DNA targets by greedy farthest-point sampling on 3-mer composition, which spreads them across the panel rather than clustering them, restricted to targets measured against at least 40% of the proteins in the panel. Additionally, I excluded from consideration any target that also appears in `dna_pwm`, `mitomi` or `rogers200`, since a target held out of one panel would otherwise reappear in another dataset’s training rows. I then selected the ten held-out proteins among those measured

against at least four of the five held-out targets, taking the ten with the largest range of measured intensity across them, in order to ensure sufficient dynamic range in the ground truth output.

4.3 Modeling

Objective and Metrics

I trained the model on groups of DNA sequences drawn from a single protein, with the objective that the model rank the relative affinities of the DNA sequences in the group. This follows from §4.2, since the PWM-derived labels only support within-protein comparisons.

Because the model is a re-ranker, I evaluate it with rank statistics rather than classification statistics. I report the Spearman correlation between predicted and measured values, computed per protein and averaged, and the enrichment of true binders in the top 10% of the predicted ranking over their base rate.

A Model over AlphaFold3 Pair Embeddings

The model takes as input the AlphaFold3 pair embedding of dimension 128, cropped to 192 tokens. This pair representation encodes learned geometric relationships between every pair of tokens in the structure. The crop keeps every DNA token and fills the remainder with the protein tokens of highest predicted contact probability against any DNA token, using the contact probabilities AlphaFold3 emits alongside the structure.

The network projects the cropped tensor to 48 channels and applies one triangle attention block, consisting of an outgoing and an incoming triangle multiplicative update followed

by row-wise and column-wise gated attention over four heads. Each of these four sublayers is output-gated with its gate initialized closed, such that the block begins training near the identity. The block output is normalized, mean-pooled over the token pairs that form the protein-DNA interface, and passed through a multilayer perceptron of widths 48, 24 and 1 (Figure 34). The model has 62,161 trainable parameters.

I trained with a differentiable relaxation of the Spearman correlation. Each training example is a group of $K = 4$ DNA sequences drawn from one protein and one assay, and each step takes $G = 2$ such groups, giving 8 complexes per gradient update. Writing s_i for the score the model assigns to the i th member of a group, the soft rank of that member is

$$\tilde{r}_i = \sum_{j=1}^K \sigma\left(\frac{s_i - s_j}{\tau}\right), \quad \sigma(x) = \frac{1}{1 + e^{-x}},$$

at temperature $\tau = 1$. Each term of the sum is a soft indicator that s_i exceeds s_j , approaching 1 as $s_i - s_j$ grows and 0 as it falls, and the $j = i$ term contributes a constant $\frac{1}{2}$ to every member, which leaves the correlation below unchanged. As $\tau \rightarrow 0$ each term approaches an exact indicator and $\tilde{r}_i$ approaches the number of group members that s_i outcores, which is its integer rank. For any $\tau > 0$, however, $\tilde{r}_i$ is a smooth function of the scores and therefore differentiable.

Writing r_i for the integer rank of the measured label of the same member, held fixed during the backward pass, the loss on a batch is

$$\mathcal{L} = 1 - \frac{1}{G} \sum_{g=1}^G \rho\left(\tilde{\mathbf{r}}^{(g)}, \mathbf{r}^{(g)}\right),$$

where ρ denotes the Pearson correlation taken over the K members of group g . The Spearman correlation between two vectors is the Pearson correlation of their ranks, so $\rho(\tilde{\mathbf{r}}, \mathbf{r})$ is the Spearman correlation with the model’s ranks softened, and minimizing $\mathcal{L}$ maximizes the quantity the model is evaluated on. The labels are oriented such that a lower value corresponds to a stronger binder.

Within a group, sequences are drawn in proportion to their percentile of measured score, which samples the stronger binders more often. I trained for 100,000 steps with AdamW at a constant learning rate of 3e-4 on two L40 GPUs.

I also trained a second model on protein-ligand and protein-protein binding data alongside the protein-DNA data. These add 1.88 million and 143,000 rows respectively, bringing the training set from 213,000 rows to 2.23 million, and batches are drawn 50% protein-DNA, 25% protein-ligand and 25% protein-protein under a fixed quota, since the ligand pool would otherwise dominate every batch. The premise was that the chemistry of a binding interface is shared across the three modalities, such that a network over interface pair representations could be trained on far more interfaces than the protein-DNA data alone contains. Both models were trained under the same split and the same recipe, and differ only in the modality mix of their batches. I refer to them below as the DNA-only model and the multimodal model.

4.4 *In Silico* Performance

Training Behaviour

I tracked validation during training on the PWM-derived data, which is in distribution, and on the three held-out quadrants of each in-house panel (Figure 35). The in-

distribution correlation plateaus around 0.4 — however, the held-out quadrants oscillate for the whole run without converging. In particular, the doubly held-out quadrant oscillates heavily — it contains only 50 and 45 protein-DNA pairs over ten proteins each, so a handful of proteins changing rank order will produce such swings.

The doubly-novel validation set is too small to be used to compare checkpoints against each other, or to draw meaningful conclusions about the relative performance between models. Given the limitations of the size of the dataset, we instead then turn to prospective evaluation on a fully held-out test set to choose between models.

Prospective Evaluation

I evaluated both models on an orthogonality assay of 48 designs that did not exist when they were trained, each measured against a panel of 16 DNA targets (Figure 36). The designs span four designed targets, and no design shares more than 30% sequence identity with any protein in the training set. Performance is the Spearman correlation between predicted and measured intensity for one design across the panel, averaged over the 48 designs (Table 5).

Both AlphaFold3 confidence metrics have no signal on the full panel, where the DNA-only model reaches 0.46 with a median per-design correlation of 0.50 and a standard error of 0.035 over the 48 designs. On the five DNA targets withheld from training the DNA-only model reaches 0.60 against 0.25 for minPAE. This is the slice that describes the intended application, as neither the protein nor the DNA was available to the model during training.

Multimodal training lowered performance on both slices. The two models differ only in the modality mix of their training batches, and the DNA-only model scores higher by 0.16 on the full panel and by 0.36 on the held-out DNA targets.

Deployment

Given the limited data, including more designs is likely to improve the model. Therefore, the production model was trained on all the available DNA binding data, including the held-out quadrants of the two in-house panels (§4.2). Although this means we do not have access to a validation set to monitor possible overfitting, we surmise that a model trained on more datapoints than a previous model to the same number of steps will not overfit. The final production model is therefore trained for 100,000 steps with AdamW, on all the available DNA binding data. It has been deployed within the lab and we are awaiting results on its utility as a prospective filter of designs.

Limitations

The project is currently limited by the availability of design data for training and validation. Ten held-out proteins and five held-out DNA sequences per in-house dataset is a small test set. The prospective assay is larger, at 48 designs — however its doubly held-out slice still only contains 5 DNA sequences per protein.

The model takes AlphaFold3 predictions as input, so its accuracy is bounded by the ability of AlphaFold3 to generate a reasonable structure prediction — a systematic error in how AlphaFold3 models a class of interface will propagate to the bind/nobind model.

The model is a re-ranker over designs that have already passed AlphaFold3 filtering.

The reported performance is therefore conditional on that filter having been applied, and does not describe unfiltered designs.

4.5 Further Work

I expect the largest improvement to come from better negatives. Currently, most training negatives are sequences that score low against a position weight matrix, such as randomly generated sequences, which are easy to separate from binders and therefore contribute little gradient late in training. Additionally, every protein was balanced to an even split of binders and non-binders when its negatives were generated (§4.2), which fixes the number of negatives per protein while leaving their difficulty uncontrolled. We could instead sample hard negatives, and over-sample the highest-scoring non-binding sequences for each protein, which would concentrate training on the near-threshold discriminations the model has to make at inference.

The multimodal result of §4.4 leaves the transfer question open. I trained both models at a fixed quota of 50% protein-DNA, 25% protein-ligand and 25% protein-protein throughout, and did not sweep it. Protein-ligand and protein-protein interfaces might still transfer under a different mixing ratio, or under a staged schedule such as pretraining on the other two modalities before a protein-DNA fine-tuning stage.

Finally, more measured data would reduce the reliance on generated labels. MITOMI and the in-house orthogonality panels give continuous measured values, and further in-house campaigns produce more of exactly the design-against-panel data the model is evaluated on.

References

- [1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O’Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, May 2024.
- [2] Arjun K. Aditham, Craig J. Markin, Daniel A. Mokhtari, Nicole DelRosso, and Polly M. Fordyce. High-Throughput Affinity Measurements of Transcription Factor and DNA Mutations Reveal Affinity and Specificity Determinants. *Cell Systems*, 12(2):112–127.e11, February 2021.
- [3] Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O’Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The Rosetta All-Atom Energy

Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, May 2017.

- [4] Han Altae-Tran, Soumya Kannan, Anthony J. Suberski, Kepler S. Mears, F. Esra Demircioglu, Lukas Moeller, Selin Kocalar, Rachel Oshiro, Kira S. Makarova, Rhiannon K. Macrae, Eugene V. Koonin, and Feng Zhang. Uncovering the functional diversity of rare CRISPR-Cas systems with deep terascale clustering. *Science*, 382(6673), November 2023.
- [5] Namrata Anand and Tudor Achim. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models, 2022.
- [6] Susan E. Andrew, Y. Paul Goldberg, Berry Kremer, Håkan Telenius, Jane Theilmann, Shelin Adam, Elizabeth Starr, Ferdinando Squitieri, Biaoyang Lin, Michael A. Kalchman, Rona K. Graham, and Michael R. Hayden. The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington’s disease. *Nature Genetics*, 4(4):398–403, August 1993.
- [7] Christian B. Anfinsen. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230, July 1973.
- [8] Ivan Anishchenko, Samuel J. Pellock, Tamuka M. Chidyausiku, Theresa A. Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K. Bera, Frank DiMaio, Lauren Carter, Cameron M. Chow, Gaetano T. Montelione, and David Baker. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, December 2021.

- [9] Andrew V. Anzalone, Luke W. Koblan, and David R. Liu. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nature Biotechnology*, 38(7):824–844, June 2020.
- [10] Justin Ashworth, James J. Havranek, Carlos M. Duarte, Django Sussman, Raymond J. Monnat, Barry L. Stoddard, and David Baker. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*, 441(7093):656–659, June 2006.
- [11] Justin Ashworth, Gregory K. Taylor, James J. Havranek, S. Arshiya Quadri, Barry L. Stoddard, and David Baker. Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic Acids Research*, 38(16):5601–5608, April 2010.
- [12] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, August 2021.
- [13] Minkyung Baek, Ryan McHugh, Ivan Anishchenko, Hanlun Jiang, David Baker,

and Frank DiMaio. Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nature Methods*, 21(1):117–121, November 2023.

- [14] Lesa J. Beamer and Carl O. Pabo. Refined 1.8 Å crystal structure of the λ repressor-operator complex. *Journal of Molecular Biology*, 227(1):177–196, September 1992.
- [15] Michael F Berger, Anthony A Philippakis, Aaron M Qureshi, Fangxue S He, Preston W Estep, and Martha L Bulyk. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology*, 24(11):1429–1435, September 2006.
- [16] Lavi S Bigman and Yaakov Levy. Proteins: molecules defined by their trade-offs. *Current Opinion in Structural Biology*, 60:50–56, February 2020.
- [17] Adam J. Bogdanove and Daniel F. Voytas. TAL Effectors: Customizable Proteins for DNA Targeting. *Science*, 333(6051):1843–1846, September 2011.
- [18] Jasper Butcher, Rohith Krishna, Raktim Mitra, Rafael I. Brent, Yanjing Li, Nathaniel Corley, Paul T. Kim, Jonathan Funk, Simon Mathis, Saman Salike, Aiko Muraishi, Helen Eisenach, Tuscan Rock Thompson, Jie Chen, Yuliya Politanska, Enisha Sehgal, Brian Coventry, Odin Zhang, Bo Qiang, Kieran Didi, Max Kazman, Frank DiMaio, and David Baker. De novo Design of All-atom Biomolecular Interactions with RFdiffusion3. *bioRxiv*, September 2025.
- [19] Longxing Cao, Brian Coventry, Inna Goreshnik, Buwei Huang, William Sheffler, Joon Sung Park, Kevin M. Jude, Iva Marković, Rameshwar U. Kadam, Koen H. G. Verschueren, Kenneth Verstraete, Scott Thomas Russell Walsh, Nathaniel

- Bennett, Ashish Phal, Aerin Yang, Lisa Kozodoy, Michelle DeWitt, Lora Picton, Lauren Miller, Eva-Maria Strauch, Nicholas D. DeBouver, Allison Pires, Asim K. Bera, Samer Halabiya, Bradley Hammerson, Wei Yang, Steffen Bernard, Lance Stewart, Ian A. Wilson, Hannele Ruohola-Baker, Joseph Schlessinger, Sangwon Lee, Savvas N. Savvides, K. Christopher Garcia, and David Baker. Design of protein-binding proteins from the target structure alone. *Nature*, 605(7910):551–560, March 2022.
- [20] Jaime A Castro-Mondragon, Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, Paul Boddie, Aziz Khan, Nicolás Manosalva Pérez, Oriol Fornes, Tiffany Y Leung, Alejandro Aguirre, Fayrouz Hammal, Daniel Schmelter, Damir Baranasic, Benoit Ballester, Albin Sandelin, Boris Lenhard, Klaas Vandepoele, Wyeth W Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 50(D1):D165–D173, November 2021.
- [21] Chai Discovery team, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis, Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, October 2024.
- [22] Brett Chevalier, Monique Turmel, Claude Lemieux, Raymond J. Monnat, and Barry L. Stoddard. Flexible DNA Target Site Recognition by Divergent Homing Endonuclease Isoschizomers I-CreI and I-MsoI. *Journal of Molecular Biology*, 329(2):253–269, May 2003.

- [23] Tsu-Pei Chiu, Satyanarayan Rao, and Remo Rohs. Physicochemical models of protein–DNA binding with standard and modified base pairs. *Proceedings of the National Academy of Sciences*, 120(4):e2205796120, January 2023.
- [24] Yehlin Cho, Griffin Rangel, Gaurav Bhardwaj, and Sergey Ovchinnikov. Protein Hunter: exploiting structure hallucination within diffusion for protein design. *bioRxiv*, 2025. bioRxiv preprint 2025-10.
- [25] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, October 2022.
- [26] Justas Dauparas, Gyu Rie Lee, Robert Pecoraro, Linna An, Ivan Anishchenko, Cameron Glasscock, and D. Baker. Atomic context-conditioned protein sequence design using LigandMPNN. *bioRxiv*, December 2023.
- [27] Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis, 2021.
- [28] Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael M. Bronstein, Martin Steinegger, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Scaling Atomistic Protein Binder Design with Generative Pretraining and Test-Time Compute. In *International Conference on Learning Representations (ICLR)*, 2026. Introduces Proteina-Complexa.

- [29] Thomas E. Ellenberger, Christopher J. Brandl, Kevin Struhl, and Stephen C. Harrison. The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted α helices: Crystal structure of the protein-DNA complex. *Cell*, 71(7):1223–1237, December 1992.
- [30] Monica Elrod-Erickson, Timothy E Benson, and Carl O Pabo. High-resolution structures of variant Zif268–DNA complexes: implications for understanding zinc finger–DNA recognition. *Structure*, 6(4):451–464, April 1998.
- [31] Minchao Fang, Chentong Wang, Jungang Shi, Fengbai Lian, Qihan Jin, Zhe Wang, Yanzhe Zhang, Peipei Chen, Zhanyuan Cui, Yanjun Wang, et al. HalluDesign: protein optimization and de novo design via iterative structure hallucination and sequence design. *bioRxiv*, 2025. bioRxiv preprint 2025-11.
- [32] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1), July 2022.
- [33] Sarel J. Fleishman and David Baker. Role of the Biomolecular Energy Gap in Protein Design, Structure, and Evolution. *Cell*, 149(2):262–273, April 2012.
- [34] Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks, 2020.
- [35] Marcel Geertz, David Shore, and Sebastian J. Maerkl. Massively parallel measurements of molecular interaction kinetics on a microfluidic platform. *Proceedings of the National Academy of Sciences*, 109(41):16540–16545, September 2012.
- [36] Tomas Geffner, Kieran Didi, Zhonglin Cao, Danny Reidenbach, Zuobai Zhang,

- Christian Dallago, Emine Kucukbenli, Karsten Kreis, and Arash Vahdat. La-Proteina: Atomistic Protein Generation via Partially Latent Flow Matching. In *International Conference on Learning Representations (ICLR)*, 2026.
- [37] Cameron J. Glasscock, Robert J. Pecoraro, Ryan McHugh, Lindsey A. Doyle, Wei Chen, Olivier Boivin, Beau Lonnquist, Emily Na, Yuliya Politanska, Hugh K. Haddox, David Cox, Christoffer Norn, Brian Coventry, Inna Goreshnik, Dionne Vafeados, Gyu Rie Lee, Raluca Gordân, Barry L. Stoddard, Frank DiMaio, and David Baker. Computational design of sequence-specific DNA-binding proteins. *Nature Structural & Molecular Biology*, 32(11):2252–2261, September 2025.
- [38] Matthias Glögl, Aditya Krishnakumar, Robert J. Ragotte, Inna Goreshnik, Brian Coventry, Asim K. Bera, Alex Kang, Emily Joyce, Green Ahn, Buwei Huang, Wei Yang, Wei Chen, Mariana Garcia Sanchez, Brian Koepnick, and David Baker. Target-conditioned diffusion generates potent TNFR superfamily antagonists and agonists. *Science*, 386(6726):1154–1161, December 2024.
- [39] Lucas B. Harrington, David Burstein, Janice S. Chen, David Paez-Espino, Enbo Ma, Isaac P. Witte, Joshua C. Cofsky, Nikos C. Kyrpides, Jillian F. Banfield, and Jennifer A. Doudna. Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science*, 362(6416):839–842, November 2018.
- [40] Stephen C. Harrison and Aneel K. Aggarwal. DNA RECOGNITION BY PROTEINS WITH THE HELIX-TURN-HELIX MOTIF. *Annual Review of Biochemistry*, 59(1):933–969, June 1990.
- [41] Charis L. Himeda, Takako I. Jones, and Peter L. Jones. Targeted epigenetic

repression by CRISPR/dSaCas9 suppresses pathogenic DUX4-fl expression in FSHD. *Molecular Therapy - Methods & Clinical Development*, 20:298–311, March 2021.

- [42] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models, 2020.
- [43] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance, 2022.
- [44] Po-Ssu Huang, Scott E. Boyken, and David Baker. The coming of age of de novo protein design. *Nature*, 537(7620):320–327, September 2016.
- [45] Maxwell A. Hume, Luis A. Barrera, Stephen S. Gisselbrecht, and Martha L. Bulyk. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic Acids Research*, 43(D1):D117–D122, November 2015.
- [46] John Ingraham, Vikas K. Garg, Regina Barzilay, and Tommi Jaakkola. Generative Models for Graph-Based Protein Design. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- [47] John B. Ingraham, Max Baranov, Zak Costello, Karl W. Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M. Lord, Christopher Ng-Thow-Hing, Erik R. Van Vlack, Shan Tie, Vincent Xue, Sarah C. Cowles, Alan Leung, João V. Rodrigues, Claudio L. Morales-Perez, Alex M. Ayoub, Robin Green, Katherine Puentes, Frank Oplinger, Nishant V. Panwar, Fritz Obermeyer, Adam R. Root, Andrew L. Beam, Frank J. Poelwijk, and Gevorg Grigoryan. Illuminating protein

space with a programmable generative model. *Nature*, 623(7989):1070–1078, November 2023.

- [48] Mark Isalan, Aaron Klug, and Yen Choo. A rapid, generally applicable method to engineer zinc fingers illustrated by targeting the HIV-1 promoter. *Nature Biotechnology*, 19(7):656–660, July 2001.
- [49] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A. Doudna, and Emmanuelle Charpentier. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science*, 337(6096):816–821, August 2012.
- [50] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael J. L. Townshend, and Ron Dror. Learning from Protein Structure with Geometric Vector Perceptrons, 2020.
- [51] Arttu Jolma, Jian Yan, Thomas Whittington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M. Vaquerizas, Renaud Vincentelli, Nicholas M. Luscombe, Timothy R. Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. DNA-Binding Specificities of Human Transcription Factors. *Cell*, 152(1-2):327–339, January 2013.
- [52] J. Keith Joung and Jeffry D. Sander. TALENs: a widely applicable technology for targeted genome editing. *Nature Reviews Molecular Cell Biology*, 14(1):49–55, November 2012.

- [53] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, July 2021.
- [54] Ahmad S. Khalil, Timothy K. Lu, Caleb J. Bashor, Cherie L. Ramirez, Nora C. Pyenson, J. Keith Joung, and James J. Collins. A Synthetic Biology Framework for Programming Eukaryotic Transcription Functions. *Cell*, 150(3):647–658, August 2012.
- [55] Shagufta H. Khan and Raj Kumar. An Overview of the Importance of Conformational Flexibility in Gene Regulation by the Transcription Factors. *Journal of Biophysics*, 2009:1–9, February 2009.
- [56] Joseph L. Kim, Dimitar B. Nikolov, and Stephen K. Burley. Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature*, 365(6446):520–527, October 1993.
- [57] Aaron Klug. The Discovery of Zinc Fingers and Their Applications in Gene Regulation and Genome Manipulation. *Annual Review of Biochemistry*, 79(1):213–231, June 2010.

- [58] Alexis C. Komor, Yongjoo B. Kim, Michael S. Packer, John A. Zuris, and David R. Liu. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, 533(7603):420–424, April 2016.
- [59] Tanja Kortemme. De novo protein design—From new structures to programmable functions. *Cell*, 187(3):526–544, February 2024.
- [60] Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S. Morey-Burrows, Ivan Anishchenko, Ian R. Humphreys, Ryan McHugh, Dionne Vafeados, Xinting Li, George A. Sutherland, Andrew Hitchcock, C. Neil Hunter, Alex Kang, Evans Brackenbrough, Asim K. Bera, Minkyung Baek, Frank DiMaio, and David Baker. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science*, 384(6693), April 2024.
- [61] Andrew Kubaney, Andrew Favor, Lilian McHugh, Raktim Mitra, Robert Pecoraro, Justas Dauparas, Cameron Glasscock, and David Baker. RNA sequence design and protein–DNA specificity prediction with NA-MPNN. *bioRxiv*, 2025. bioRxiv preprint 2025-10.
- [62] Brian Kuhlman, Gautam Dantas, Gregory C. Ireton, Gabriele Varani, Barry L. Stoddard, and David Baker. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science*, 302(5649):1364–1368, November 2003.
- [63] Anna Lauko, Samuel J. Pellock, Kiera H. Sumida, Ivan Anishchenko, David Jurgens, Woody Ahern, Jihun Jeung, Alexander F. Shida, Andrew Hunt, Indrek Kal-

- vet, et al. Computational design of serine hydrolases. *Science*, 388(6744):eadu2454, 2025.
- [64] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023.
- [65] Sidney Lyayuga Lisanza, Jacob Merle Gershon, Samuel W. K. Tipps, Jeremiah Nelson Sims, Lucas Arnoldt, Samuel J. Hendel, Miriam K. Simma, Ge Liu, Muna Yase, Hongwei Wu, Claire D. Tharp, Xinting Li, Alex Kang, Evans Brackenbrough, Asim K. Bera, Stacey Gerben, Bruce J. Wittmann, Andrew C. McShan, and David Baker. Multistate and functional protein design using RoseTTAFold sequence space diffusion. *Nature Biotechnology*, 43(8):1288–1298, September 2024.
- [66] X.-J. Lu. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31(17):5108–5121, September 2003.
- [67] Xiang-Jun Lu, Harmen J. Bussemaker, and Wilma K. Olson. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Research*, 43(21):e142, 2015.
- [68] N. M. Luscombe. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Research*, 29(13):2860–2874, July 2001.

- [69] Nicholas M Luscombe, Susan E Austin, Helen M Berman, and Janet M Thornton.
An overview of the structures of protein-DNA complexes. *Genome Biology*,
1(1):reviews001.1, 2000.
- [70] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand,
Raphael R. Eguchi, Po-Ssu Huang, and Richard Socher. ProGen: Language
Modeling for Protein Generation, 2020.
- [71] Sanaa Mansoor, Minkyung Baek, Hahnbeom Park, Gyu Rie Lee, and David Baker.
Protein Ensemble Generation Through Variational Autoencoder Latent Space
Sampling. *Journal of Chemical Theory and Computation*, 20(7):2689–2695, March
2024.
- [72] Leslie Milk, Robert Daber, and Mitchell Lewis. Functional rules for lac repres-
sor-operator associations and implications for protein-DNA interactions. *Protein
Science*, 19(6):1162–1172, May 2010.
- [73] Raktim Mitra, Ari S Cohen, Jared M Sagendorf, Helen M Berman, and Remo Rohs.
DNAProDB: an updated database for the automated and interactive analysis of
protein-DNA complexes. *Nucleic Acids Research*, 53(D1):D396–D402, November
2024.
- [74] Raktim Mitra, Jinsan Li, Jared M. Sagendorf, Yibei Jiang, Ari S. Cohen, Tsu-
Pei Chiu, Cameron J. Glasscock, and Remo Rohs. Geometric deep learning of
protein-DNA binding specificity. *Nature Methods*, 21(9):1674–1683, 2024.
- [75] R K Moyzis, J M Buckingham, L S Cram, M Dani, L L Deaven, M D Jones,

- J Meyne, R L Ratliff, and J R Wu. A highly conserved repetitive DNA sequence, (TTAGGG)_n, present at the telomeres of human chromosomes. *Proceedings of the National Academy of Sciences*, 85(18):6622–6626, September 1988.
- [76] Monica E. Neugebauer, Alvin Hsu, Mandana Arbab, Nicholas A. Krasnow, Amber N. McElroy, Smriti Pandey, Jordan L. Doman, Tony P. Huang, Aditya Ragu-ram, Samagya Banskota, Gregory A. Newby, Jakub Tolar, Mark J. Osborn, and David R. Liu. Evolution of an adenine base editor into a small, efficient cytosine base editor with low off-target activity. *Nature Biotechnology*, 41(5):673–685, May 2023.
- [77] Edwin N. Neumann, Tessa M. Bertozzi, Elaine Wu, Fiona Serack, John W. Harvey, Pamela P. Brauer, Catherine P. Pirtle, Alissa Coffey, Michael Howard, Nikita Kamath, Kenney Lenz, Kenia Guzman, Michael H. Raymond, Ahmad S. Khalil, Benjamin E. Deverman, Eric Vallabh Minikel, Sonia M. Vallabh, and Jonathan S. Weissman. Brainwide silencing of prion protein by AAV-mediated delivery of an engineered compact epigenetic editor. *Science*, 384(6703):eado7082, June 2024.
- [78] Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brix, et al. Sequence modeling and design from molecular to genome scale with Evo. *Science*, 386(6723):eado9336, 2024.
- [79] Wilma K. Olson, Yun Li, and Marcia O. Fenley. Insights into DNA solvation found in protein-DNA structures. *Biophysical Journal*, 121(24):4749–4758, December 2022.

- [80] Z. Otwinowski, R. W. Schevitz, R.-G. Zhang, C. L. Lawson, A. Joachimiak, R. Q. Marmorstein, B. F. Luisi, and P. B. Sigler. Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, 335(6188):321–329, September 1988.
- [81] Raphaël Pantier, Kashyap Chhatbar, Grace Alston, Heng Yang Lee, and Adrian Bird. High-throughput sequencing SELEX for the determination of DNA-binding protein specificities in vitro. *STAR Protocols*, 3(3):101490, September 2022.
- [82] Julia M. Rogers, Luis A. Barrera, Deepak Reyon, Jeffry D. Sander, Manolis Kellis, J. Keith Joung, and Martha L. Bulyk. Context influences on TALE–DNA binding revealed by quantitative profiling. *Nature Communications*, 6(1):7440, June 2015.
- [83] Remo Rohs, Xiangshu Jin, Sean M. West, Rohit Joshi, Barry Honig, and Richard S. Mann. Origins of Specificity in Protein-DNA Recognition. *Annual Review of Biochemistry*, 79(1):233–269, June 2010.
- [84] Remo Rohs, Sean M. West, Alona Sosinsky, Peng Liu, Richard S. Mann, and Barry Honig. The role of DNA shape in protein–DNA recognition. *Nature*, 461(7268):1248–1253, October 2009.
- [85] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, 2021.
- [86] N C Seeman, J M Rosenberg, and A Rich. Sequence-specific recognition of double helical nucleic acids by proteins. *Proceedings of the National Academy of Sciences*, 73(3):804–808, March 1976.

- [87] Enisha Sehgal, Yuliya Politanska, Raktim Mitra, Paul T. Kim, Nayim González Rodríguez, Tushar Warriar, Andrew Kubaney, Akira Morishita, Riley Quijano, Jasper Butcher, Rohith Krishna, Robert J. Pecoraro, Brian Belmont, Nicole Roullier, Inna Goreshnik, Dionne K. Vafeados, Paul Kwon, Rachana Ramarao, Jussi Taipale, Cameron J. Glasscock, and David Baker. Generative design of sequence specific DNA binding proteins. *bioRxiv*, 2026. Preprint; Chapter 2 is based on this paper. ptkim is co-first-author.
- [88] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Židek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, January 2020.
- [89] Matthew Slattery, Tianyin Zhou, Lin Yang, Ana Carolina Dantas Machado, Raluca Gordân, and Remo Rohs. Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences*, 39(9):381–399, September 2014.
- [90] Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O’Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, et al. BoltzGen: toward universal binder design. *bioRxiv*, 2025. bioRxiv preprint 2025-11.
- [91] Summer B. Thyme, Jordan Jarjour, Ryo Takeuchi, James J. Havranek, Justin Ashworth, Andrew M. Scharenberg, Barry L. Stoddard, and David Baker. Ex-

ploitation of binding energy for catalysis and design. *Nature*, 461(7268):1300–1304, October 2009.

- [92] Frederic Tremblay, Qiang Xiong, Shrijal S. Shah, Chih-Wei Ko, Kenneth Kelly, Mary S. Morrison, Cristiana Giancarlo, Ricardo N. Ramirez, Erica M. Hildebrand, Sarah B. Voytek, Gabriel K. El Sebae, Shane H. Wright, Liam Lofgren, Scott Clarkson, Christine Waters, Samantha J. Linder, Songlei Liu, Taesun Eom, Shefal Parikh, Yuki Weber, Salette Martinez, Padma Malyala, Sahar Abubucker, Ari E. Friedland, Morgan L. Maeder, Angelo Lombardo, Vic E. Myer, and Aron B. Jaffe. A potent epigenetic editor targeting human PCSK9 for durable reduction of low-density lipoprotein cholesterol levels. *Nature Medicine*, 31(4):1329–1338, February 2025.
- [93] Umut Y. Ulge, David A. Baker, and Raymond J. Monnat. Comprehensive computational design of mCreI homing endonuclease cleavage specificity for genome engineering. *Nucleic Acids Research*, 39(10):4330–4339, February 2011.
- [94] Ilya E. Vorontsov, Irina A. Eliseeva, Arsenii Zinkevich, Mikhail Nikonov, Sergey Abramov, Alexandr Boytsov, Vasily Kamenets, Alexandra Kasianova, Semyon Kolmykov, Ivan S. Yevshin, Alexander Favorov, Yulia A. Medvedeva, Arttu Jolma, Fedor Kolpakov, Vsevolod J. Makeev, and Ivan V. Kulakovskiy. HOCOMOCO in 2024: a rebuild of the curated collection of binding models for human and mouse transcription factors. *Nucleic Acids Research*, 52(D1):D154–D163, January 2024.
- [95] Russell T. Walton, Kathleen A. Christie, Madelynn N. Whittaker, and Ben-

- jamin P. Kleinstiver. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science*, 368(6488):290–296, April 2020.
- [96] Chentong Wang, Yannan Qu, Zhangzhi Peng, Yukai Wang, Hongli Zhu, Dachuan Chen, and Longxing Cao. Proteus: Exploring Protein Structure Generation for Enhanced Designability and Efficiency. *bioRxiv*, February 2024.
- [97] Joy Y. Wang and Jennifer A. Doudna. CRISPR technology: A decade of genome editing is only the beginning. *Science*, 379(6629), January 2023.
- [98] Jue Wang, Sidney Lisanza, David Juergens, Doug Tischler, Joseph L. Watson, Karla M. Castro, Robert Ragotte, Amijai Saragovi, Lukas F. Milles, Minkyung Baek, Ivan Anishchenko, Wei Yang, Derrick R. Hicks, Marc Expòsit, Thomas Schlichthaerle, Jung-Ho Chun, Justas Dauparas, Nathaniel Bennett, Basile I. M. Wicky, Andrew Muenks, Frank DiMaio, Bruno Correia, Sergey Ovchinnikov, and David Baker. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, July 2022.
- [99] J. D. WATSON and F. H. C. CRICK. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, April 1953.
- [100] Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung

Baek, and David Baker. De novo design of protein structure and function with RFDiffusion. *Nature*, 620(7976):1089–1100, July 2023.

- [101] Matthew T. Weirauch, Ally Yang, Mihai Albu, Atina G. Cote, Alejandro Montenegro-Montero, Philipp Drewe, Hamed S. Najafabadi, Samuel A. Lambert, Ishminder Mann, Kate Cook, Hong Zheng, Alejandra Goity, Harm van Bakel, Jean-Claude Lozano, Mary Galli, Mathew G. Lewsey, Eryong Huang, Tuhin Mukherjee, Xiaoting Chen, John S. Reece-Hoyes, Sridhar Govindarajan, Gad Shaulsky, Albertha J. M. Walhout, François-Yves Bouget, Gunnar Ratsch, Luis F. Larrondo, Joseph R. Ecker, and Timothy R. Hughes. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*, 158(6):1431–1443, September 2014.
- [102] Sean M. West, Remo Rohs, Richard S. Mann, and Barry Honig. Electrostatic Interactions between Arginines and the Minor Groove in the Nucleosome. *Journal of Biomolecular Structure and Dynamics*, 27(6):861–866, June 2010.
- [103] B. I. M. Wicky, L. F. Milles, A. Courbet, R. J. Ragotte, J. Dauparas, E. Kinfu, S. Tipps, R. D. Kibler, M. Baek, F. DiMaio, X. Li, L. Carter, A. Kang, H. Nguyen, A. K. Bera, and D. Baker. Hallucinating symmetric protein assemblies. *Science*, 378(6615):56–61, October 2022.
- [104] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1 Democratizing Biomolecular Interaction Modeling. *bioRxiv*, November 2024.

- [105] Scot A. Wolfe, Lena Nekludova, and Carl O. Pabo. DNA Recognition by Cys₂His₂ Zinc Finger Proteins. *Annual Review of Biophysics and Biomolecular Structure*, 29(1):183–212, June 2000.
- [106] Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, and Jian Peng. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, July 2022.
- [107] Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3):1496–1503, January 2020.
- [108] Chen Yanover and Philip Bradley. Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. *Nucleic Acids Research*, 39(11):4564–4576, February 2011.
- [109] Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. SE(3) diffusion model with application to protein backbone generation. *arXiv*, 2023.
- [110] Vinicius Zambaldi, David La, Alexander E. Chu, Harshnira Patani, Amy E. Danson, Tristan O. C. Kwan, Thomas Frerix, Rosalia G. Schneider, David Saxton, Ashok Thillaisundaram, Zachary Wu, Isabel Moraes, Oskar Lange, Eliseo Papa, Gabriella Stanton, Victor Martin, Sukhdeep Singh, Lai H. Wong, Russ Bates, Simon A. Kohl, Josh Abramson, Andrew W. Senior, Yilmaz Alguel, Mary Y. Wu, Irene M. Aspalter, Katie Bentley, David L. V. Bauer, Peter Cherepanov,

Demis Hassabis, Pushmeet Kohli, Rob Fergus, and Jue Wang. De novo design of high-affinity protein binders with AlphaProteo, 2024.

- [111] Y. Zhang. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7):2302–2309, April 2005.
- [112] Zhaonan Zou, Tazro Ohta, Takeya Kasukawa, and Shinya Oki. ChIP-Atlas 2025 update: 10-year anniversary of a data-mining platform for exploring epigenomic landscape. *Nucleic Acids Research*, 54(W1):W77–W84, April 2026.

Figures

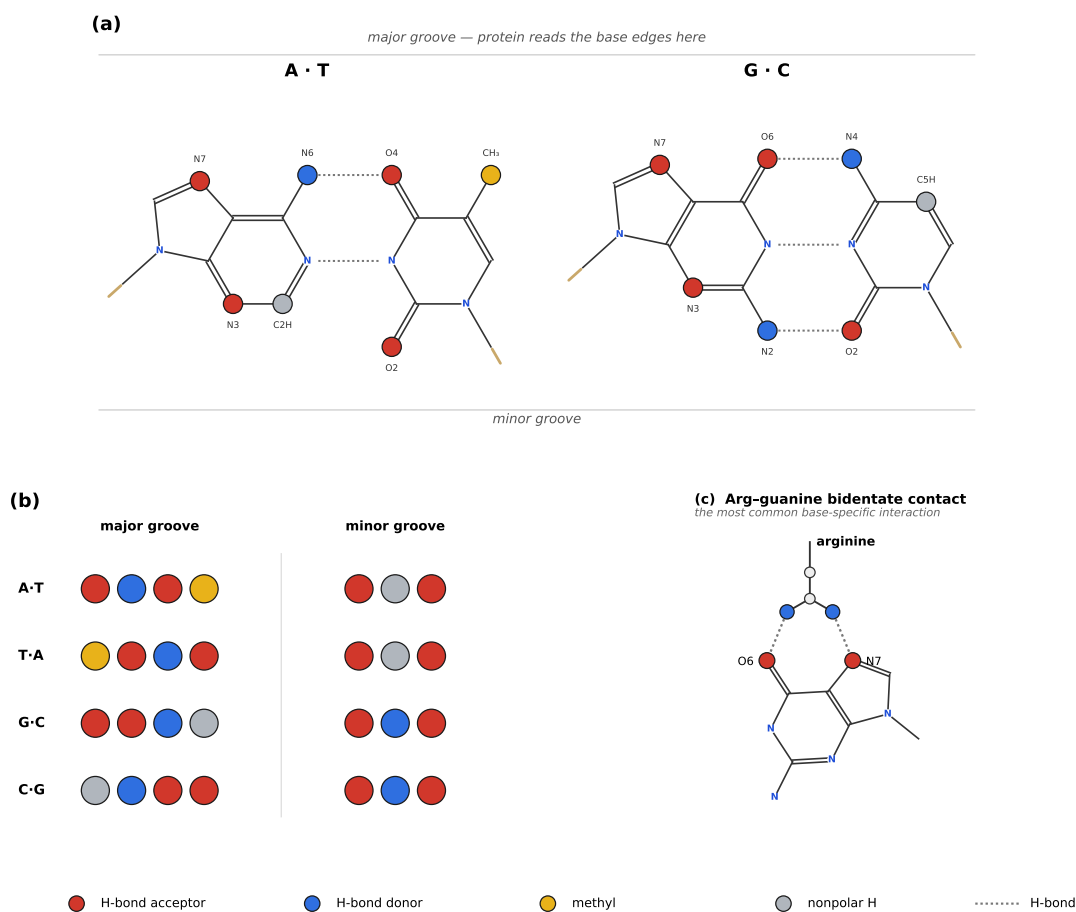


Figure 1 — Chemical patterns that distinguish base pairs at the groove edges. a, The Watson-Crick A · T and G · C pairs. b, The readout pattern of all four base pairs. Each is distinct along the major-groove edge, while along the minor-groove edge A · T matches T · A and G · C matches C · G, which is why the major groove is the more informative face for sequence readout [83]. c, The arginine-guanine bidentate contact [68].

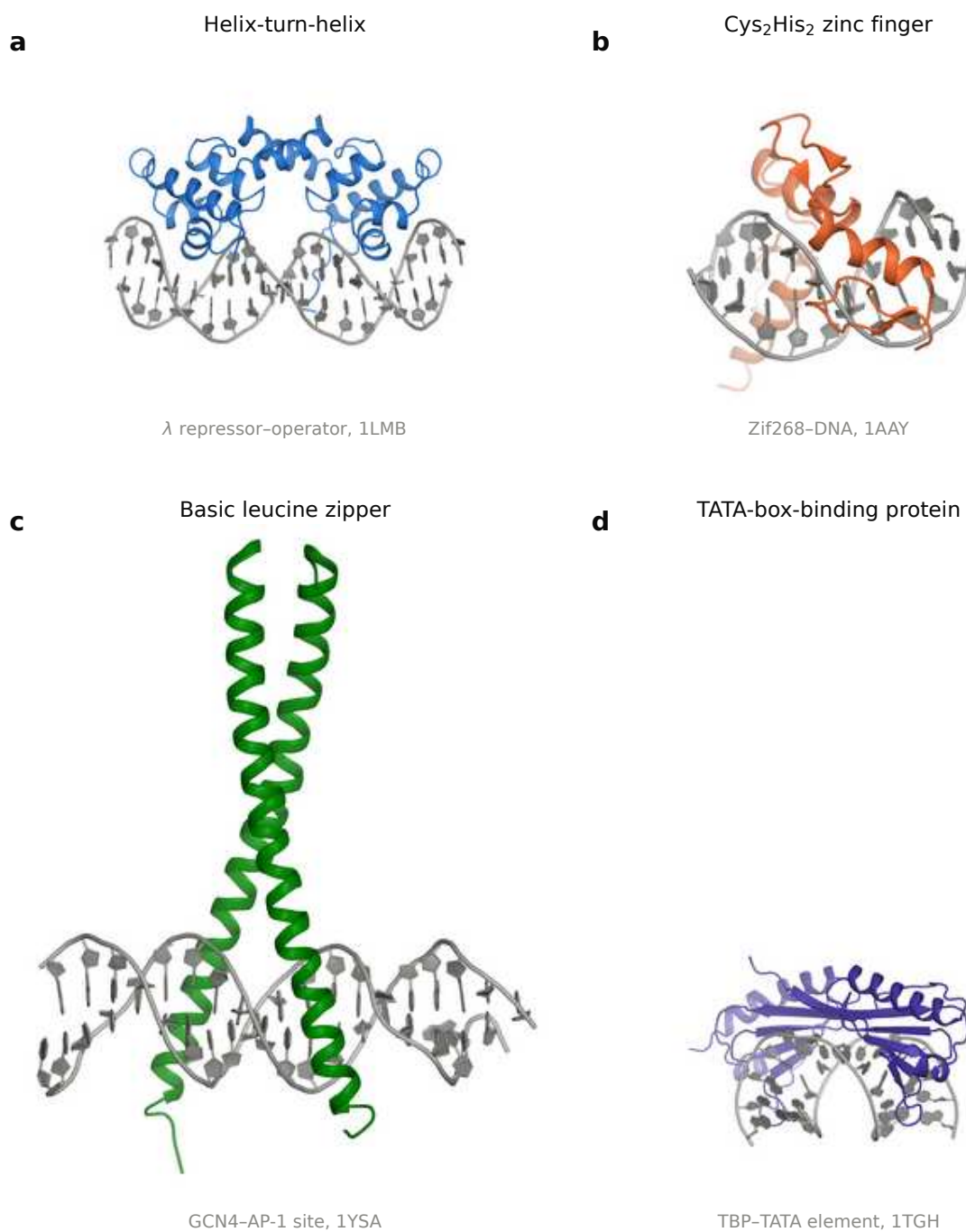


Figure 2 — Four recurring DNA-binding scaffolds. PDB entries 1LMB [14], 1AAY [30], 1YSA [29] and 1TGH [56].

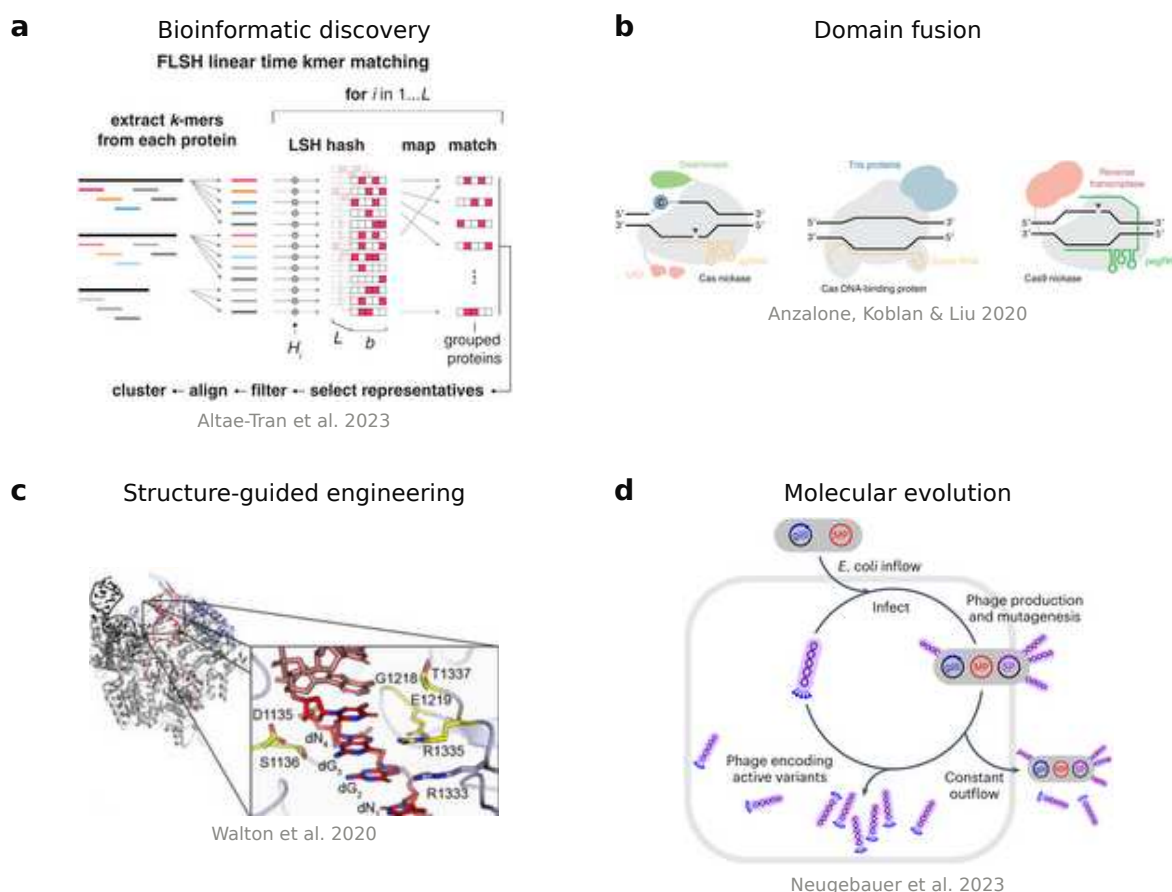


Figure 3 — Established routes to sequence-specific DNA-binding function. a, Bioinformatic discovery, here the FLSH clustering algorithm behind a terascale CRISPR-Cas survey. Adapted from Altae-Tran et al., *Science* 2023 [4]. b, Domain fusion. Adapted from Anzalone, Koblan & Liu, *Nature Biotechnology* 2020 [9]. c, Structure-guided engineering, here the residues of *Streptococcus pyogenes* Cas9 that contact the protospacer-adjacent motif, whose substitution gives the near-PAMless variants SpG and SpRY. Adapted from Walton et al., *Science* 2020 [95]. d, Molecular evolution, here phage-assisted continuous evolution. Adapted from Neugebauer et al., *Nature Biotechnology* 2023 [76].

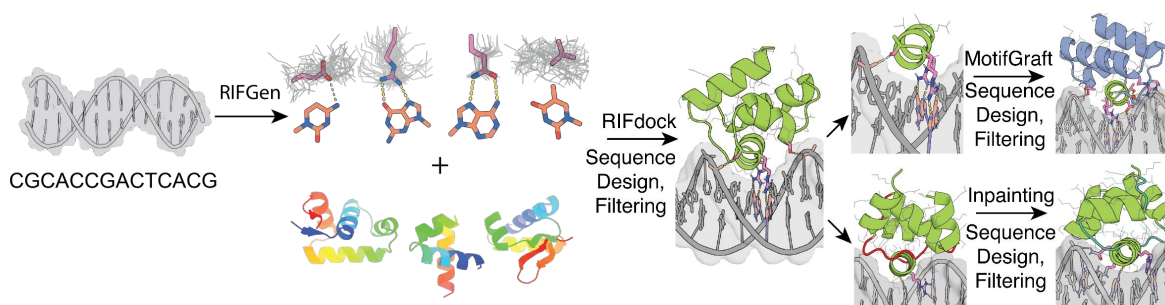


Figure 4 — The RIFdock design pipeline of Glasscock et al. RIFGen enumerates disembodied side-chain rotamers making bidentate hydrogen bonds to the major-groove base edges of the target sequence, drawn in pink against the grey cloud of alternatives sampled at each position. RIFdock then places a library of small helix-turn-helix scaffolds against that field, and MotifGraft and Inpainting elaborate the interface. Adapted from Glasscock et al. [37].

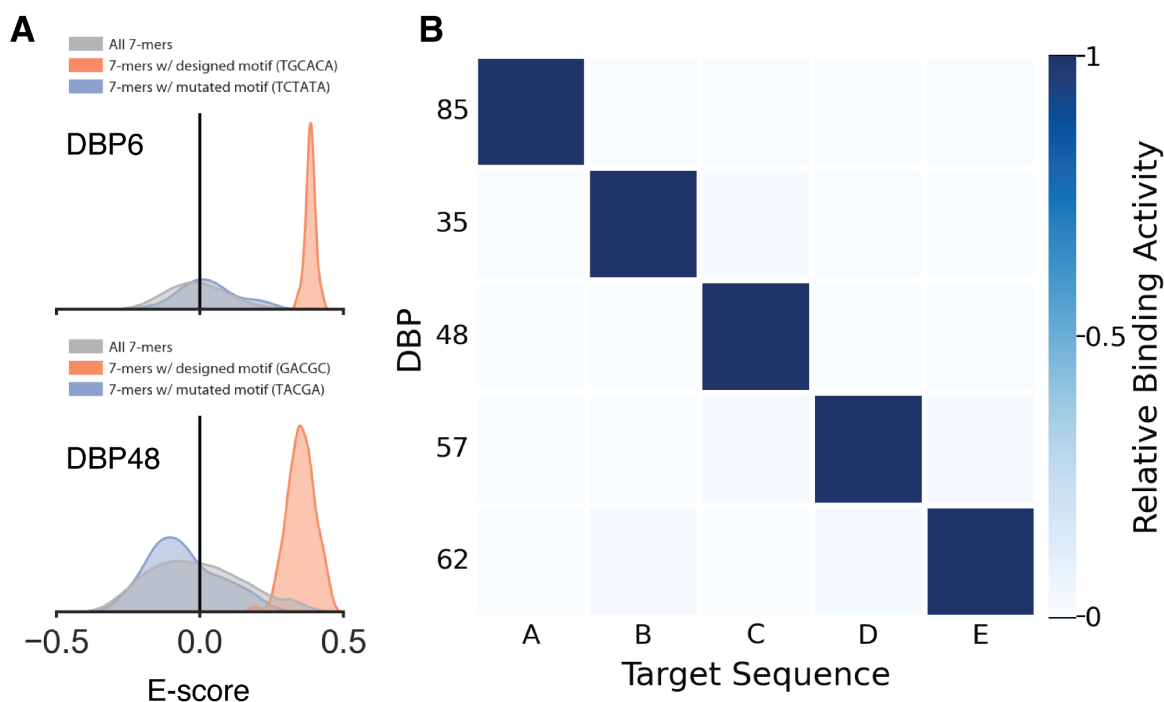


Figure 5 — Designed DNA-binding proteins of Glasscock et al. are sequence-specific. a, E-score distributions from universal protein-binding microarrays, which measure every 7-mer. b, Screened by yeast display at 1 μ M DNA with avidity, normalized by row. Adapted from Fig. 4 of Glasscock et al. [37].

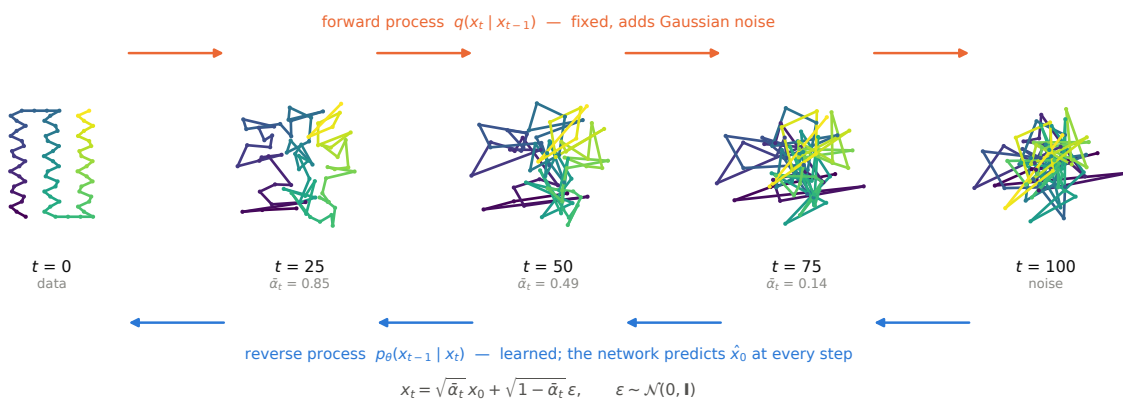


Figure 6 — The forward and reverse processes of a denoising diffusion model. An alpha-carbon trace at five points along the noising schedule. Each panel applies $x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon$ at the indicated timestep [42].

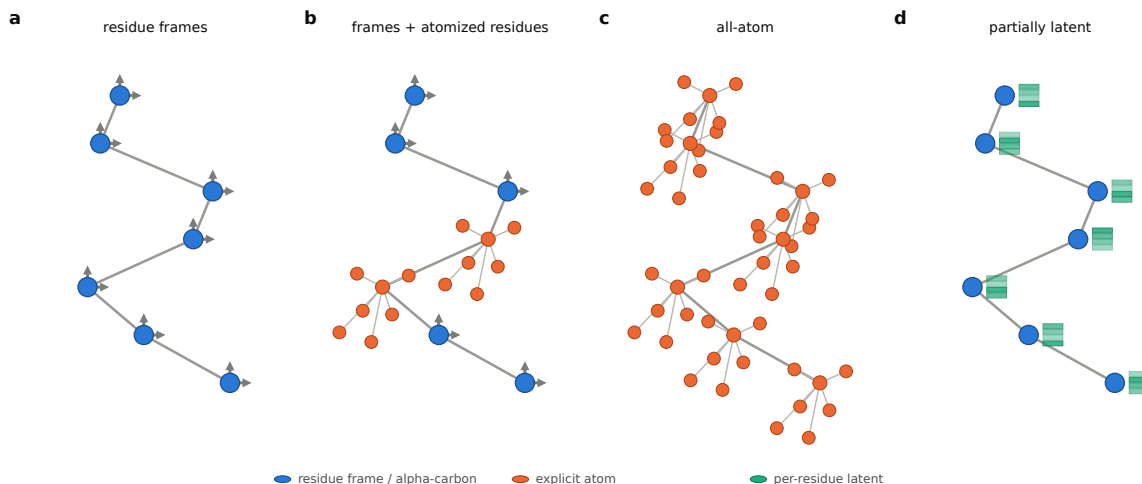


Figure 7 — Structural representations used by structure prediction and design models. a, One token per residue, with the side-chain atoms recovered from the frame by the torsion angles about the rotatable bonds. b, Chosen residues atomized into explicit atom nodes, as in RoseTTAFold All-Atom [60]. c, Every atom a separate token, as in AlphaFold3 and RFdiffusion3 [1, 18]. d, An alpha carbon per residue, with sequence and the remaining atomic detail in a per-residue latent of fixed dimension [36].

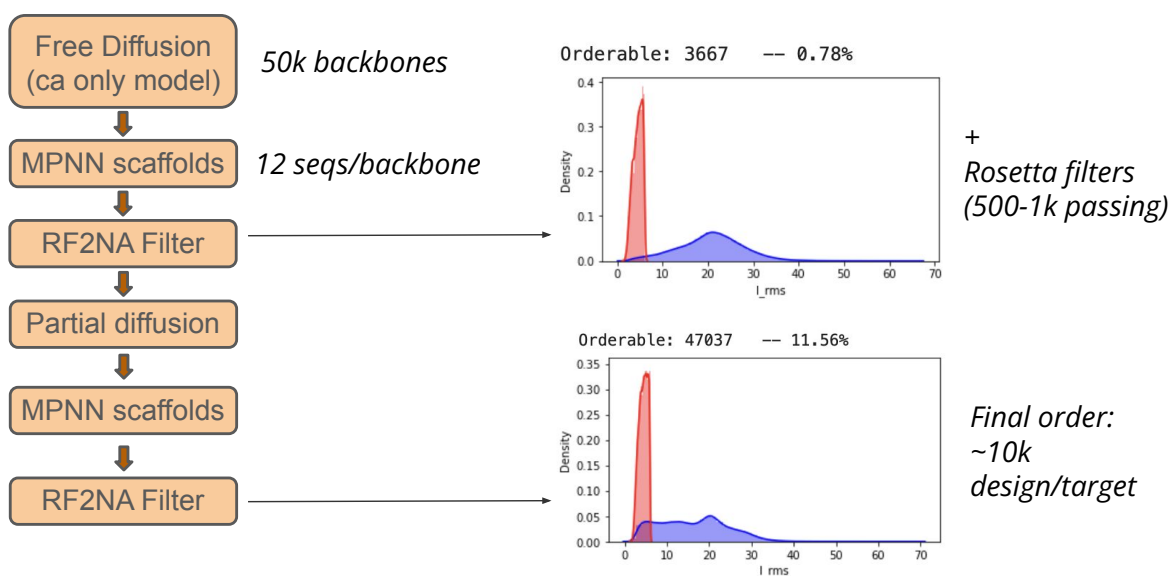


Figure 8 — The alpha-carbon diffusion design pipeline and its RoseTTAFoldNA pass rates. The stages of the pipeline are at the left, with the sampling depth at each. The plots on the right show the distributions of the RoseTTAFoldNA ligand RMSD metric in the total generated sample (blue) vs the selected designs at each stage (red).

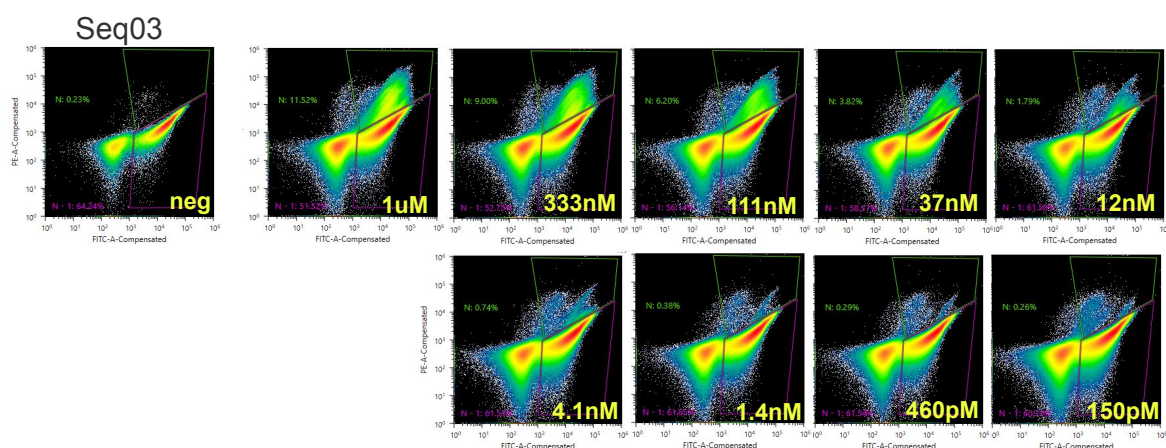


Figure 9 — Yeast-display sorting of the alpha-carbon diffusion library against one DNA target. The library was incubated with the biotinylated Seq03 DNA target and sorted on DNA binding (PE) against surface expression (FITC). The green polygon is the collection gate. Collection streaks indicate binding at sub nanomolar levels.

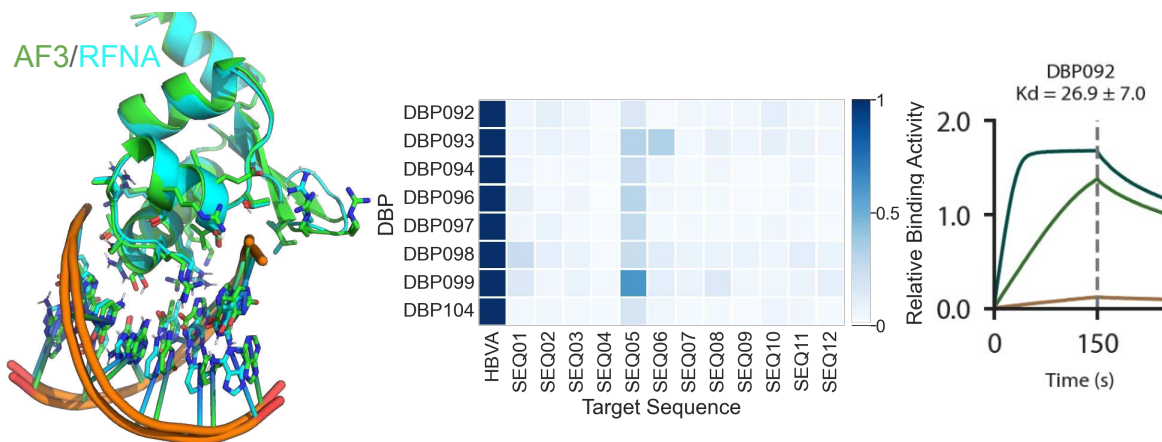


Figure 10 — Alpha-carbon diffusion designs against the HBVA target bind with high affinity and selectivity. Left, the AlphaFold3 and RoseTTAFoldNA predictions of DBP092. Middle, relative binding activity by yeast display for eight designs against HBVA and twelve other DNA targets. Right, BLI curve for DBP092, with the dissociation constant from global fitting to a 1:1 binding model.

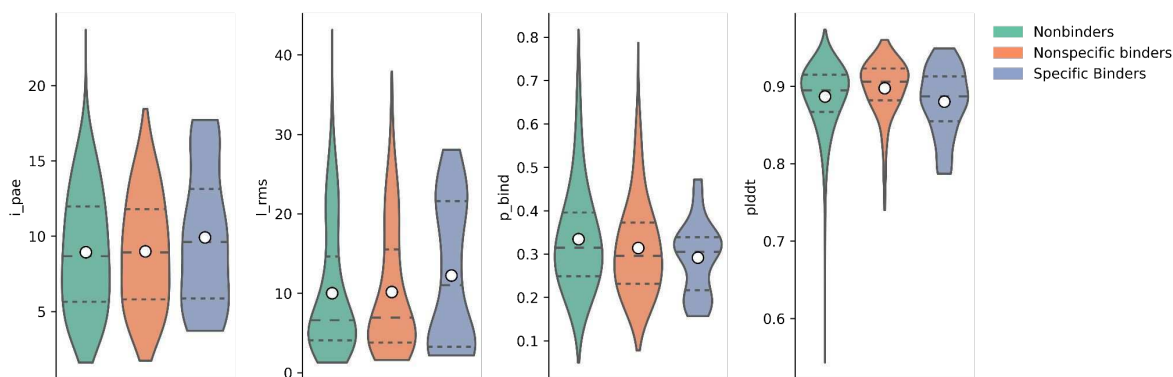


Figure 11 — RoseTTAFoldNA metrics of alpha-carbon diffusion designs, grouped by experimental outcome. White circles are group means. The experimentally determined specific binders score worse than the nonbinders according to these metrics.

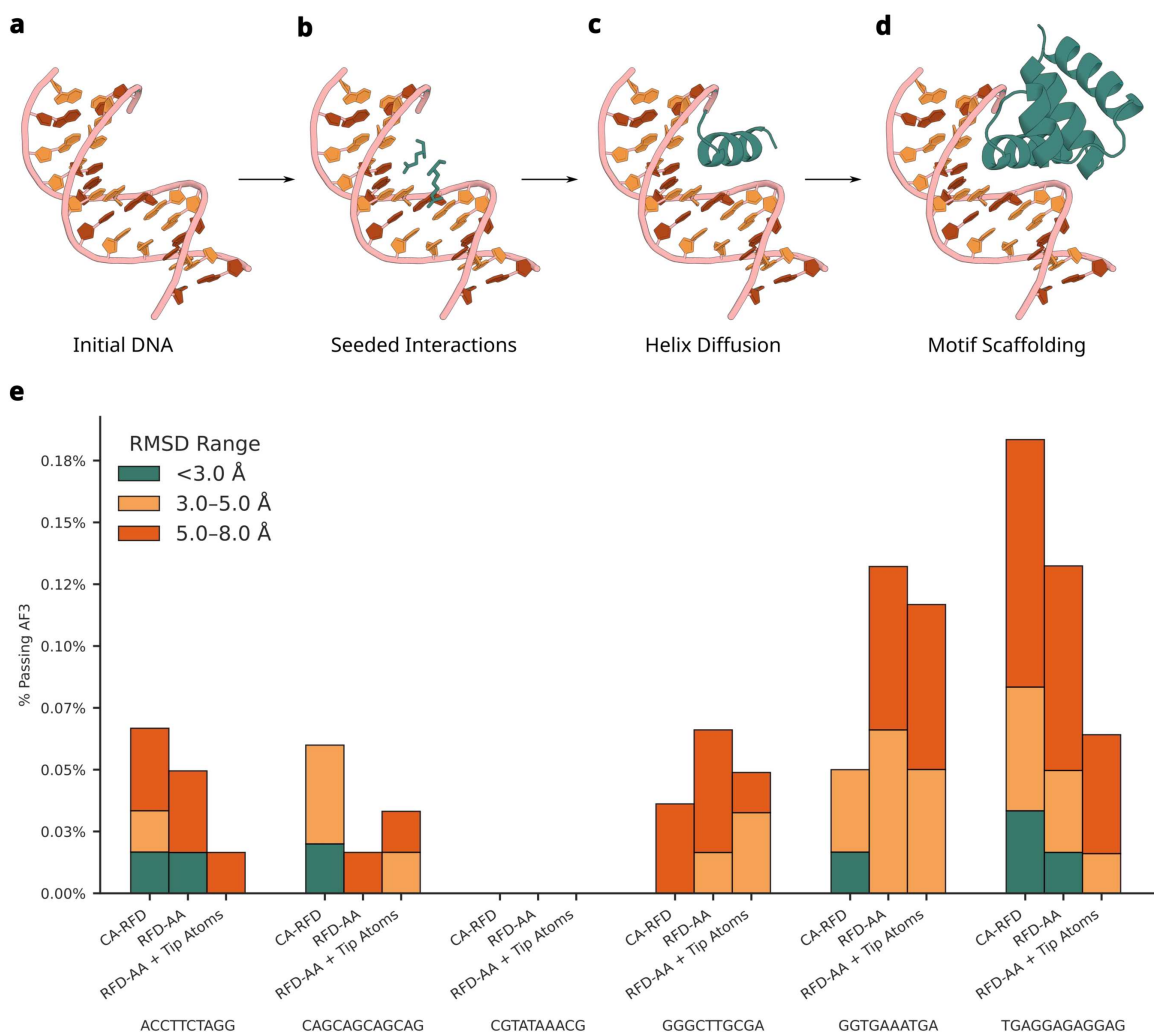


Figure 12 — The RFdiffusion All-Atom pipeline for DNA binder design, and its *in silico* pass rates. a, An initial DNA structure is generated using AlphaFold3. b, RIFGen samples side chains against the major-groove base atoms, drawn from the residues that hydrogen-bond DNA bases in native complexes in the Protein Data Bank, in order to ensure sampled interaction geometries are biophysically plausible [19]. c, The placed residue tip atoms and the DNA are used as residue and tip-atom context to run generation with a fine-tuned RFdiffusion All-Atom model, which generates a helix with residues that reproduce the specified contacts. d, A motif-scaffolding step takes high quality helices and generates a motif around them. e, Results from in-silico benchmarking of design strategies, comparing CA-RFdiffusion, RFdiffusion All-Atom, and RFdiffusion All-Atom using this tip-atom strategy, adapted from Fig. S21 of Sehgal et al. [87].

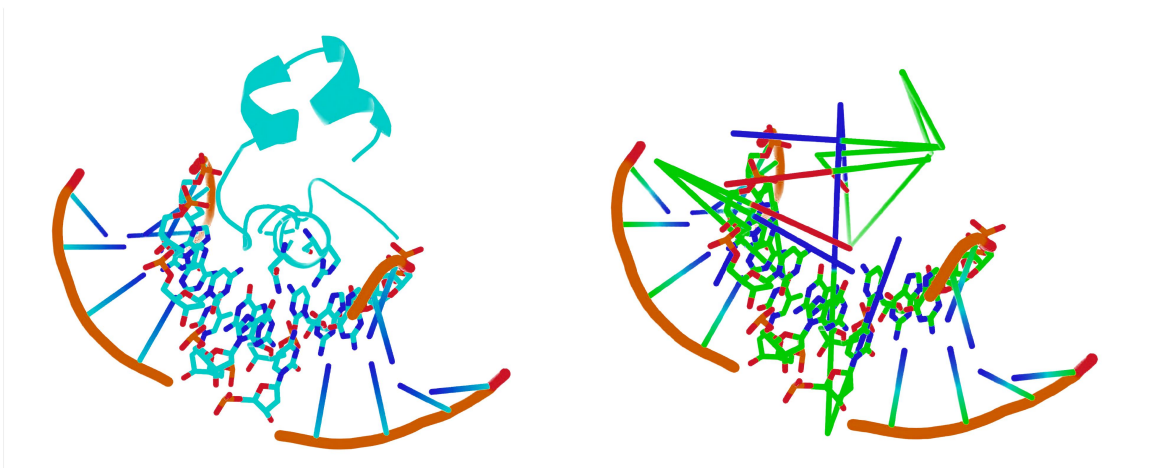


Figure 13 — Input to the tip-atom RFdiffusion All-Atom model. A designed complex (left), and what the model receives at inference (right). The DNA residues that contact the protein are atomized and fixed, the noncontacting DNA residues are provided as residues, and the guideposted tip atoms are placed at the interface. The model builds a protein that holds those contacts against the target, so the interface is specified atom by atom before any backbone exists.

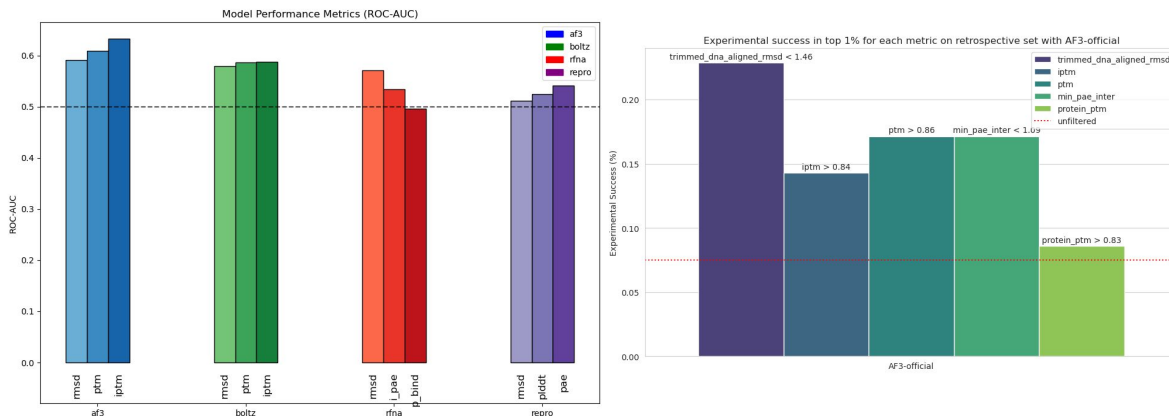


Figure 14 — Structure-prediction metrics as *in silico* filters, evaluated retrospectively. Areas under the receiver operating characteristic curve are against experimental binding, over the accumulated retrospective design set. The unfiltered experimental success rate of that set is about 0.075%, marked by the red dotted line. AlphaFold3 does enrich for binding but the enrichment rate remains low. “Repro” refers to an in-house reproduction of AlphaFold3.

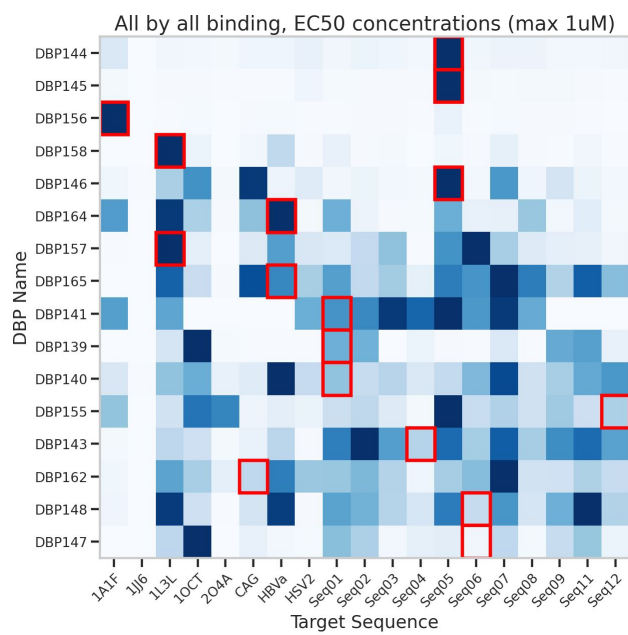
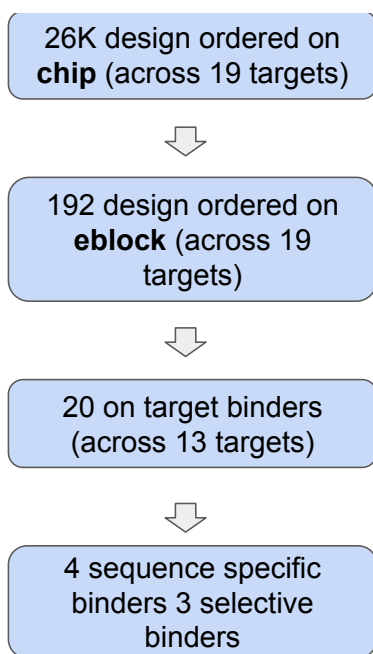


Figure 15 — Outcome of the Rfdiffusion All-Atom design campaign.

Left, the number of designs at each stage of the campaign, from the chip-scale order to the sequence-specific binders. Right, half-maximal binding concentrations measured by yeast display, with each design's intended target boxed in red. Some designs bind one target and nothing else (DBP156, DBP158), others bind half the panel (DBP141, DBP165, DBP143), and DBP147 binds the off-target 1OCT strongly while its intended Seq06 is near-white. The experimental characterization was performed by Enisha Sehgal.

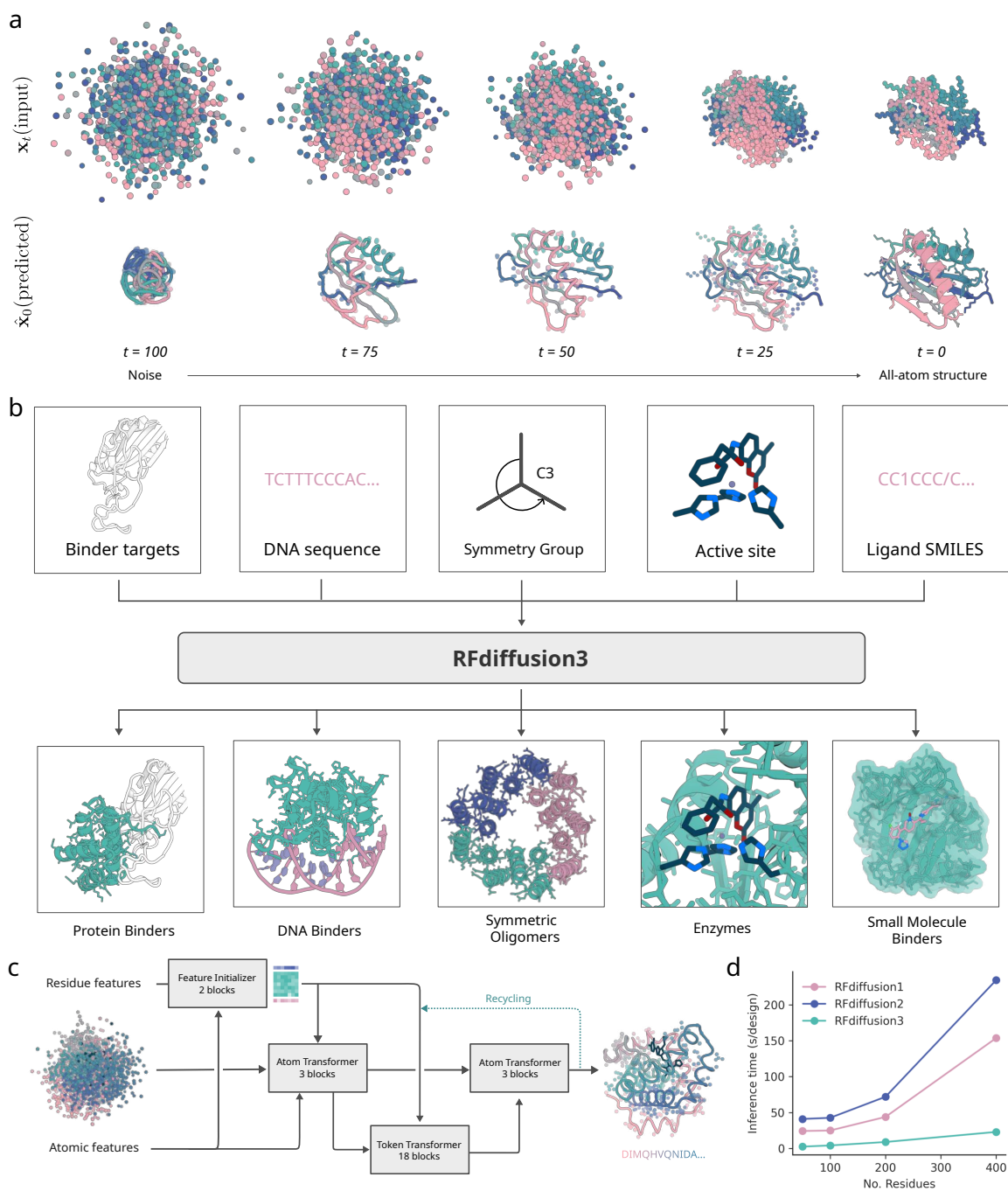


Figure 16

Figure 16 — All-atom protein design with RFdiffusion3. a, Depiction of 100 step RFD3 diffusion trajectory, predicting both backbone and sidechain atoms. (top) noisy coordinates at different timesteps. (bottom) the network prediction at each time-step. b, Generation of biomolecular interactions. (top) different conditions can be fed to the model. (bottom) sample output structures for each input condition. c, Model architecture. Arrows show information flow, grey boxes indicate modules with learnable parameters. d, Inference times for RFdiffusion variants across length scales. RFD3 is an order of magnitude faster than previous versions of RFdiffusion because of its simplified architecture (runtimes are measured on NVIDIA A6000 GPUs). Reproduced from Fig. 1 of Butcher et al. [18].

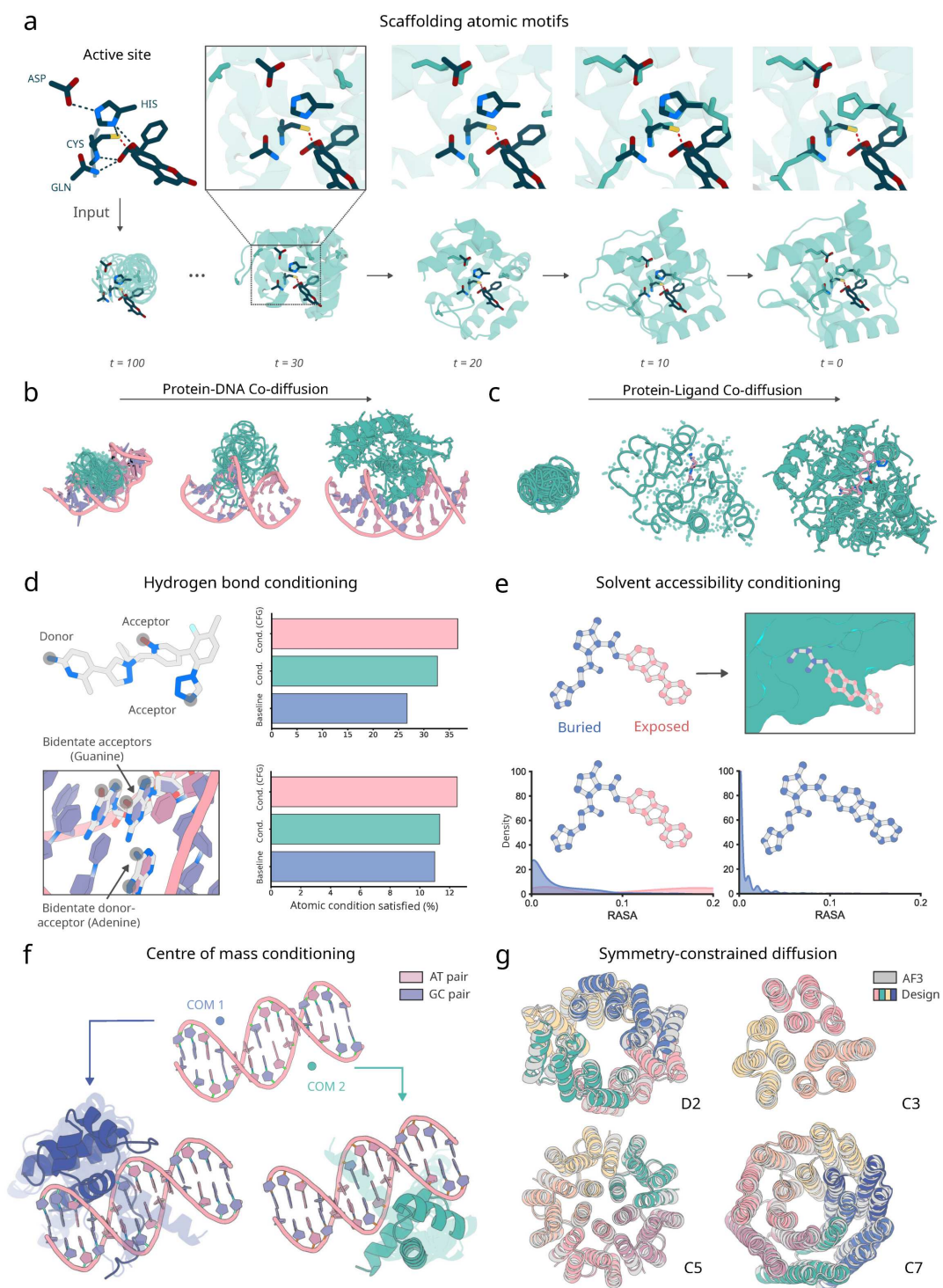


Figure 17

Figure 17 — Global and atom-level conditioning with RFdiffusion3.

a, Diffusion of enzyme active sites. Sidechains are diffused for all residues of the backbone, the optimal position in sequence to scaffold the atomic motif is thereby found during generation, allowing active sites to be scaffolded from simple active site constraints. (top) Zoom in of a trajectory showing the network deciding where in the sequence the side chain residues belong. (bottom) Visualization of predicted denoised structures across different time scales. b, Depiction of a diffusion trajectory where both the protein and the DNA are co-generated. In this case, the DNA sequence is given and only the conformation is sampled. c, Depiction of a diffusion trajectory where both the protein and the small molecule coordinates are co-generated. In this case, the ligand identity is given and only the conformation is sampled. d, Atom-level hydrogen bond conditioning. Both acceptor and donor atoms can be specified as atom level constraints for ligands (top) and nucleic acids (bottom). Applying classifier-free guidance improves adherence to the conditions in both cases. e, Atom-level conditioning of solvent accessibility by relative solvent accessible surface area (RSA). (top) example of atomic specification of accessible surface area input specification and output structure (bottom) distributions of the generated structures with different input specifications (over 400 designs from the small molecule benchmark). f, The generated centre-of-mass can be guided by the initialized noise cloud. (top) DNA input with two different protein centres of mass. (bottom) generated structures superimposed (with transparency) cluster around the input centre of mass in each case. g, Symmetric scaffolds can be specified at inference time by symmetrizing outputs of the diffusion modules. Shown are D2, C3, C5, and C7 all with AF3 C α RMSDs 0.832 Å, 0.450 Å, 0.614 Å, and 0.539 Å respectively. Reproduced from Fig. 2 of Butcher et al. [18].

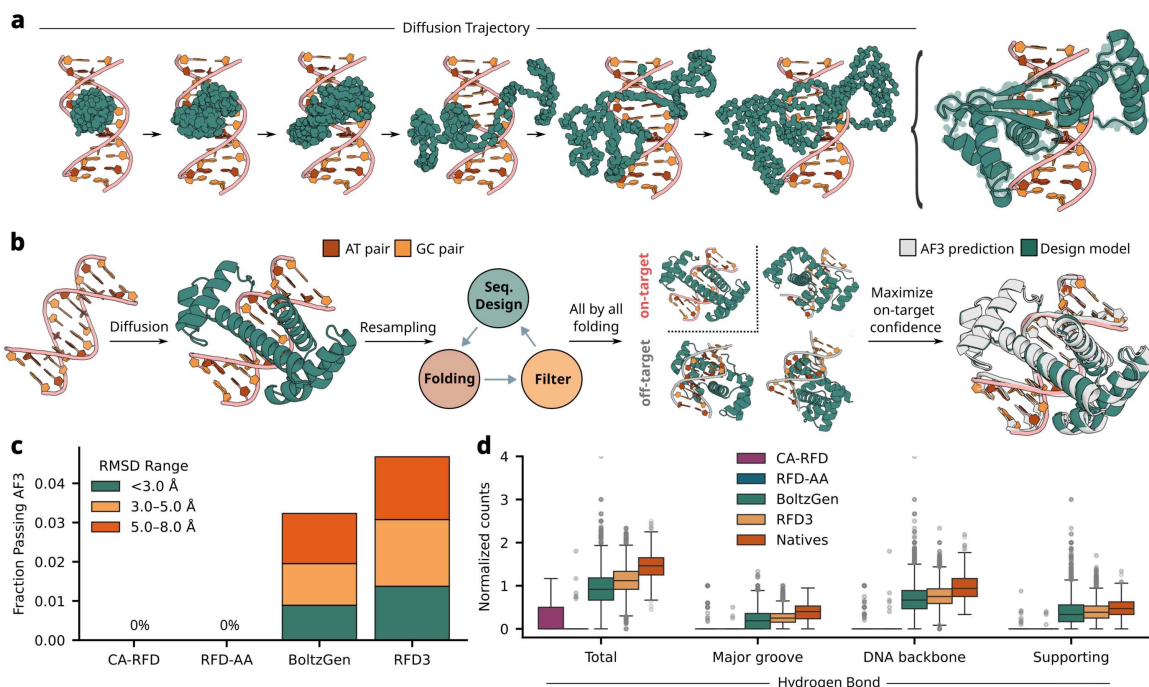


Figure 18 — Design approach. a, Snapshots of RFD3 diffusion trajectory around a target DNA structure; only protein backbone atoms shown for clarity. b, Schematic of design pipeline. RFD3 is used to generate protein structures around a specific DNA target. The design is refined by multiple rounds of LigandMPNN sequence design, AF3 prediction and filtering, improving protein-DNA co-complex geometry (binder block, Figure 19). This is followed by the specificity block (Figure 19) which selects designs for which AF3 on-target predictions have higher confidence than AF3 off-target predictions. c, AF3 DNA-aligned protein RMSD pass rates compared across different generative models. DNA binder diffusion outputs with length randomly sampled from 120-150 amino acids (1000 backbones per target per model for core 6 targets (Table 1), 5 sequences per backbone with LigandMPNN). d, Interaction counts, per interacting nucleotide, across different generative models, calculated (using DSSR [67] version 1.7.8) on AF3 predictions post sequence design with LigandMPNN. The native protein set consists of 357 transcription factor PDB structures from the JASPAR database [20] with an average information content above 1.5. Reproduced from Fig. 1 of Sehgal et al. [87].

Figure 19 — Extended depiction of the de novo DNA binder design pipeline. a, Binder block workflow. An initial AF3 prediction is generated for the target DNA sequence. Centers of mass (COMs; white spheres) are defined on the DNA and used as distinct initialization points for diffusion-based backbone generation. Hydrogen-bond donors (green arrows) and acceptors (purple arrows) on the DNA bases are annotated and can be used to guide the diffusion trajectory. Protein backbones (120–150 amino acids) are generated from these configurations using diffusion, followed by sequence design with LigandMPNN. Designed sequences are evaluated with AF3, and candidates with $\text{RMSD} < 8 \text{ \AA}$ are iteratively refined by resampling with LigandMPNN on the AF3-predicted structures. Final designs passing filtering criteria are selected for pooled synthesis and high-throughput screening. b, Specificity block workflow. Designs from the binder block with $\text{minPAE} < 1.25$ are subjected to additional rounds of sequence optimization with LigandMPNN and AF3 evaluation. Specificity is assessed through all-by-all *in silico* folding, in which each design is modeled with both on-target and off-target DNA sequences using AF3 with the protein structure templated from the latest prediction. Designs meeting structural quality thresholds ($\text{RMSD} < 1.5 \text{ \AA}$ and $\text{iPTM} > 0.9$) are further evaluated using ΔminPAE (see Figure 20). Candidates are ranked by ΔminPAE , and the top 96 designs are selected for synthesis and plate-scale experimental characterization. Reproduced from Fig. S1 of Sehgal et al. [87].

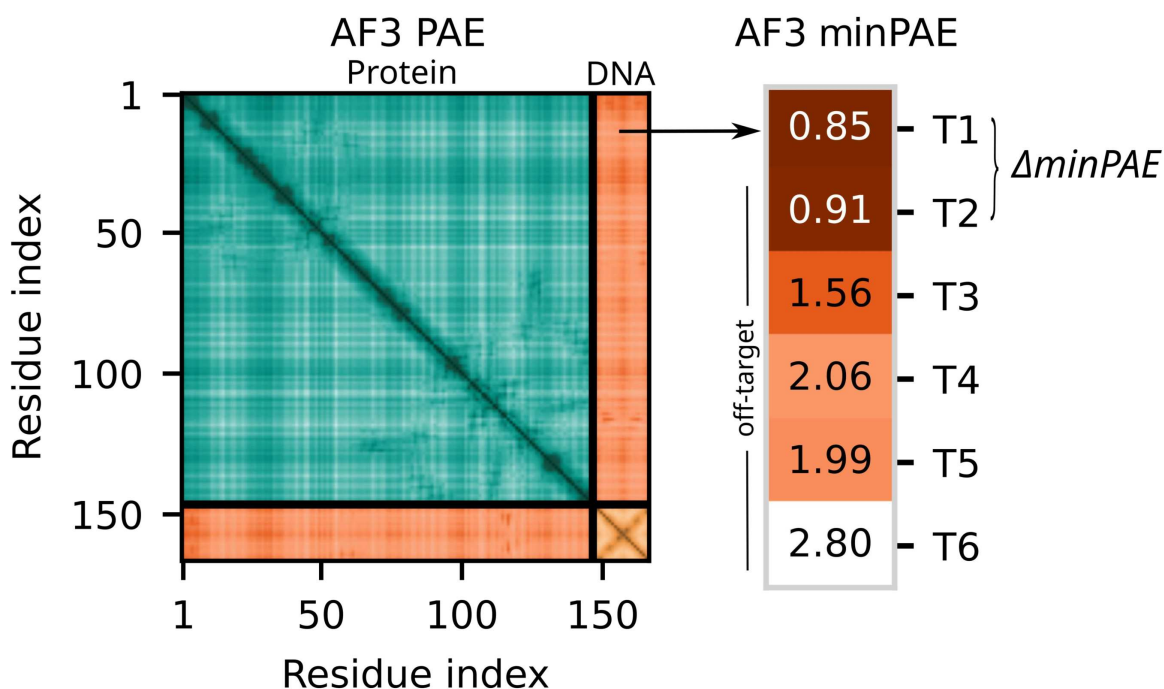


Figure 20 — Schematic description of Δ_{minPAE} calculation. AF3 predicted aligned error (PAE) values shown as a residue by residue matrix for a predicted complex: protein-protein (teal), DNA-DNA (yellow), protein-DNA (orange). Protein-DNA values are used to derive the minPAE score for a given complex (left side). The resulting minPAE value is computed for a given protein folded across all targets (T1-T6) and the difference between the on-target minPAE (T1) and the best scoring off-target minPAE values (T2-T6) is the Δ_{minPAE} (§3.2) value for the *de novo* DB under consideration (right side). Reproduced from Fig. S4 of Sehgal et al. [87].

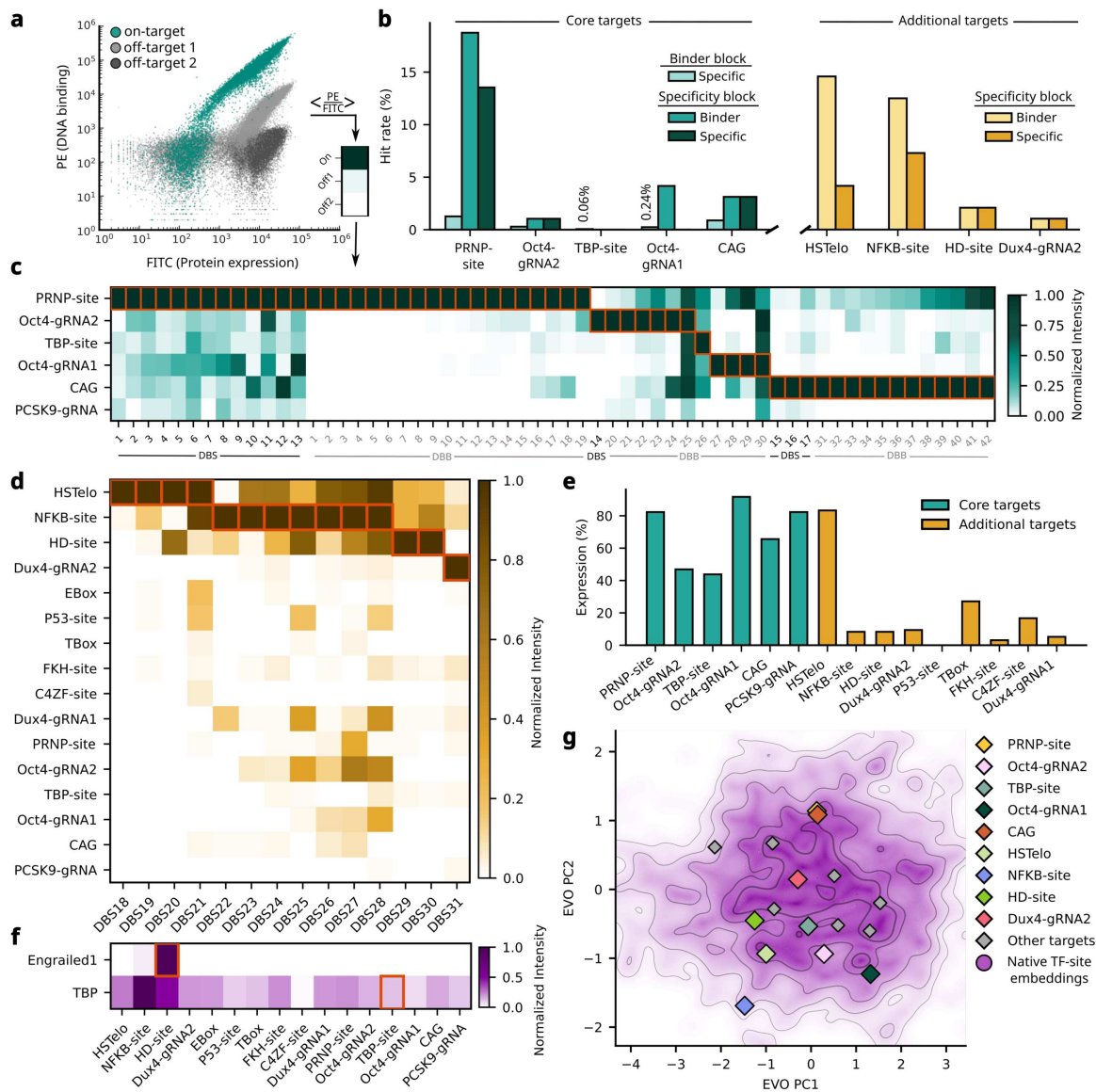


Figure 21

Figure 21 — Experimental characterization of selectivity of designed DNA binding proteins. a, Overlay of representative flow cytometry traces for a single *de novo* DNA binder (DB), illustrating binding to on-target DNA (teal) and off-target DNA sequences (light and dark gray). Mean normalized PE/FITC values for each condition are used to generate the heatmaps shown in (c,d). b, Hit rates stratified by design category. Bars show the fraction of designs meeting binding and specificity criteria, grouped by binder block and specificity block libraries. For the specificity block, the binder hit rate (medium teal, light yellow) denotes designs with detectable on-target binding, whereas the specificity hit rate (light teal, dark teal, dark yellow) denotes designs with preferential binding to the on-target sequence ($\text{PE/FITC}_{\text{on}} > \text{PE/FITC}_{\text{off}}$). Targets without a successful specific design not shown. c, All-by-all specificity matrix for DBs against six core targets (Table 1), showing PE/FITC binding signal normalized by column. Designs were screened by yeast surface display at 1 μM DNA (non-avidity format). Red boxes denote the intended on-target sequence for each design. Gray labels indicate binder block designs; black labels indicate specificity block designs. d, All-by-all specificity matrix for DBs against additional targets (Table 1), displayed as in (c). e, Expression rates of specificity block designs measured by yeast surface display. Expression was quantified via FITC signal, with a threshold of <5 -fold separation between low- and high-FITC populations used to classify expressing versus non-expressing designs. f, All-by-all specificity matrix for a subset of native transcription factors (TFs), displayed as in (c). Red boxes denote annotated target sequences derived from PWMs in the JASPAR database [20]. UniProt IDs for Engrailed1: P02836, and TBP: A0A7K4JSS6. g, Principal component analysis of Evo-derived embeddings of native TF binding sequences, with *de novo* DB target sequences projected into the same space. PC1 and PC2 explain 63.2% and 19.8% of the variance, respectively. Reproduced from Fig. 2 of Sehgal et al. [87].

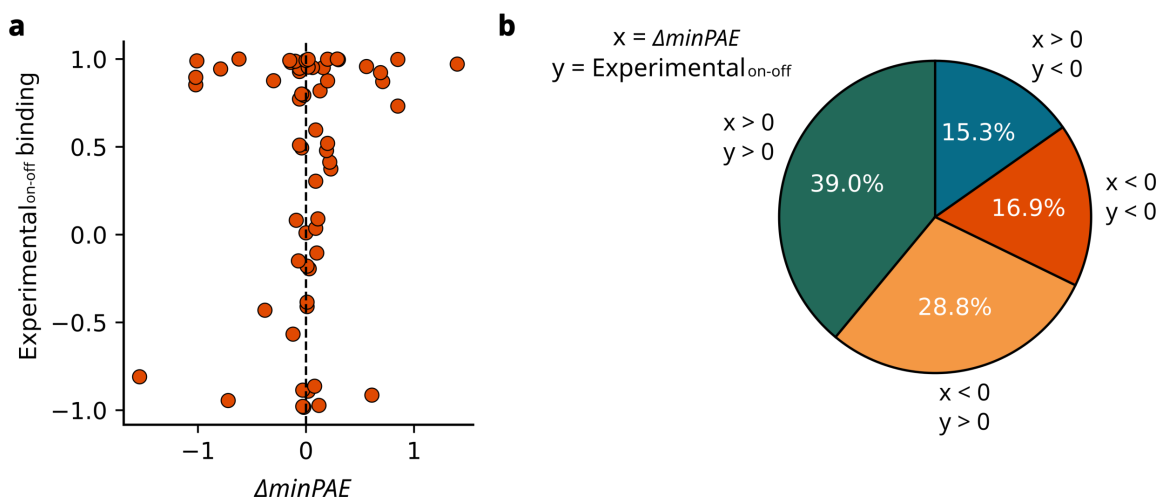


Figure 22 — Correlation between experimental binding and ΔminPAE . a, Scatter plot of ΔminPAE versus experimental binding specificity for individual designs. Experimental specificity is defined as the difference between on-target and strongest off-target normalized binding signal ($\text{PE}/\text{FITC}_{\text{on}} - \max(\text{PE}/\text{FITC}_{\text{off}})$). Analysis was performed on the binder block dataset hits ($n = 59$). b, Distribution of designs by agreement between computational and experimental specificity. Pie chart showing the fraction of designs for which ΔminPAE and experimental specificity are both positive, both negative, or discordant (opposite signs). Among designs with $\Delta\text{minPAE} > 0$ (the applied selection criterion), 71.8% (23 of 32) exhibit experimentally validated on-target selectivity. Reproduced from Fig. S7 of Sehgal et al. [87].

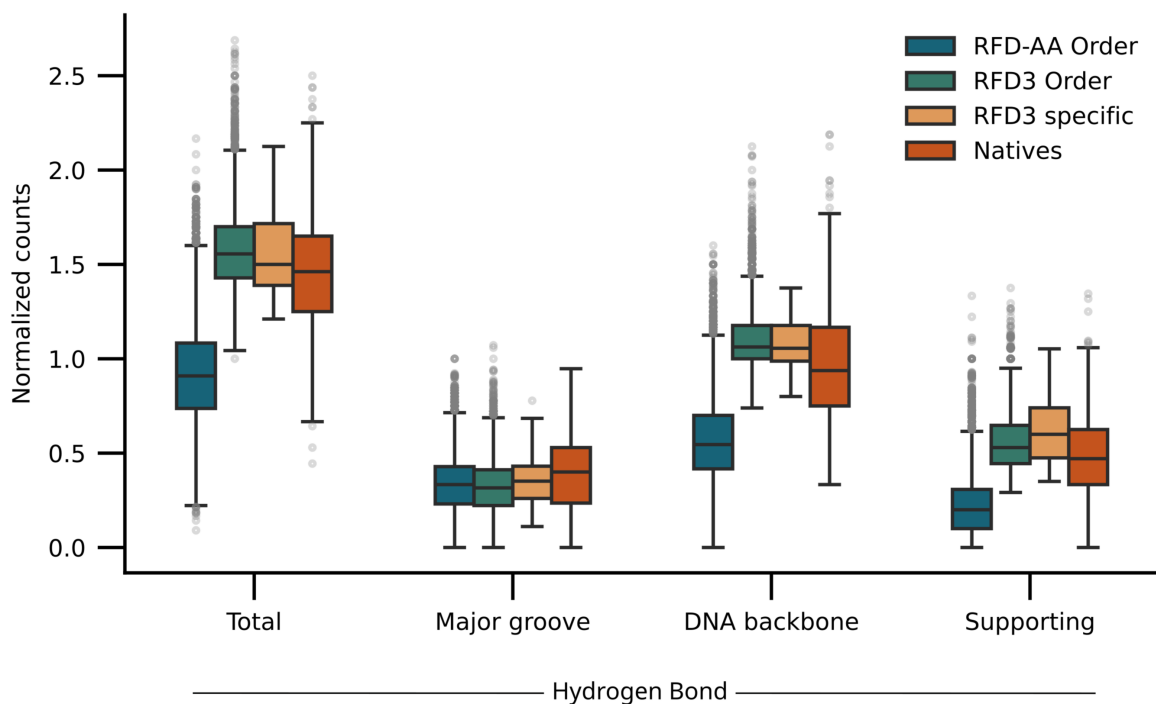


Figure 23 — Biophysical comparison of designed and native protein-DNA interfaces. Per-nucleotide interaction counts for protein-DNA complexes across design campaigns and native transcription factors (TFs). Interactions were computed from AF3 predicted structures using DSSR [67] (v1.7.8). The RFD-AA set comprises 26,575 designs, generated using the RFD-AA pipeline described in Figure 12. The RFD3 set comprises 12,653 designs, and the RFD3-specific subset includes 56 designs exhibiting sequence selective binding. Native TFs include 357 structures from the JASPAR database [20] with mean information content >1.5 . Reproduced from Fig. S8 of Sehgal et al. [87].

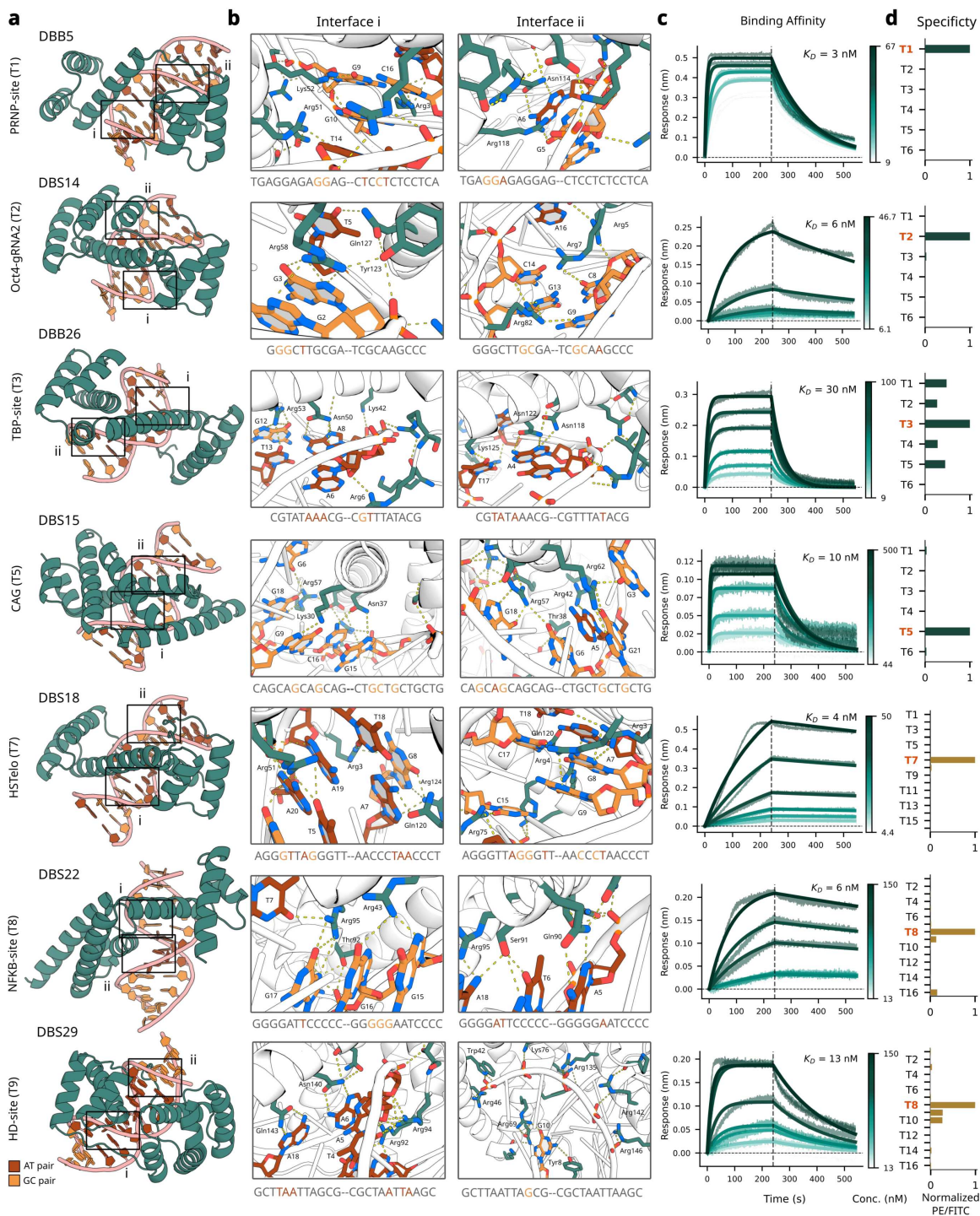


Figure 24

Figure 24 — Structures and on-target binding affinities of selected specific designs. a, Designed protein-DNA complexes. Two interaction regions (interfaces i and ii) are indicated with boxes and shown in detail in (b). b, Close-up views of protein-DNA interfaces. DNA bases are colored by identity (A/T, red; G/C, orange), and key protein side chains are shown in teal. The full DNA target sequence is indicated below each panel, with interacting bases highlighted according to the same color scheme. Selected interactions are labeled. Base numbers in the reverse complement strand are offset by target length. c, Bio-layer interferometry (BLI) binding kinetics for selected designs. Sensorgrams show association and dissociation across a titration series of protein concentrations. Equilibrium dissociation constants (K_D) were determined by global fitting to a 1:1 binding model using Octet Analysis Studio. d, Specificity profiles for selected designs, showing normalized binding signal (PE/FITC) across DNA targets. The intended on-target sequence for each design is indicated in orange on the y-axis (Table 1 for target annotations). Teal bars represent hits from the core target set and yellow represents the hits for the additional target set. Reproduced from Fig. 3 of Sehgal et al. [87].

Figure 25 — Characterization of specificity at single base resolution. a, Design models (AF3 predictions) of selected designs characterized for fine-grained specificity using yeast surface display-based competition assay, assessing binding towards all possible single nucleotide variants of corresponding DNA target. b, Protein-DNA contacts of selected regions shown per design. Nucleotides with no non-phosphate interaction are in white. Positions are annotated as per panel (c). c, Results of competition assay (normalized PE/FITC) for each design (orange-green heatmap) presented in context of average AF3 pLDDT of contacting protein residues (within 4 Å) per base-pair position (purple heatmap). Green indicates a mutant competitor target was weaker than the intended (WT) base at that position, while orange indicates higher binding for the mutant competitor. Reproduced from Fig. 4 of Sehgal et al. [87].

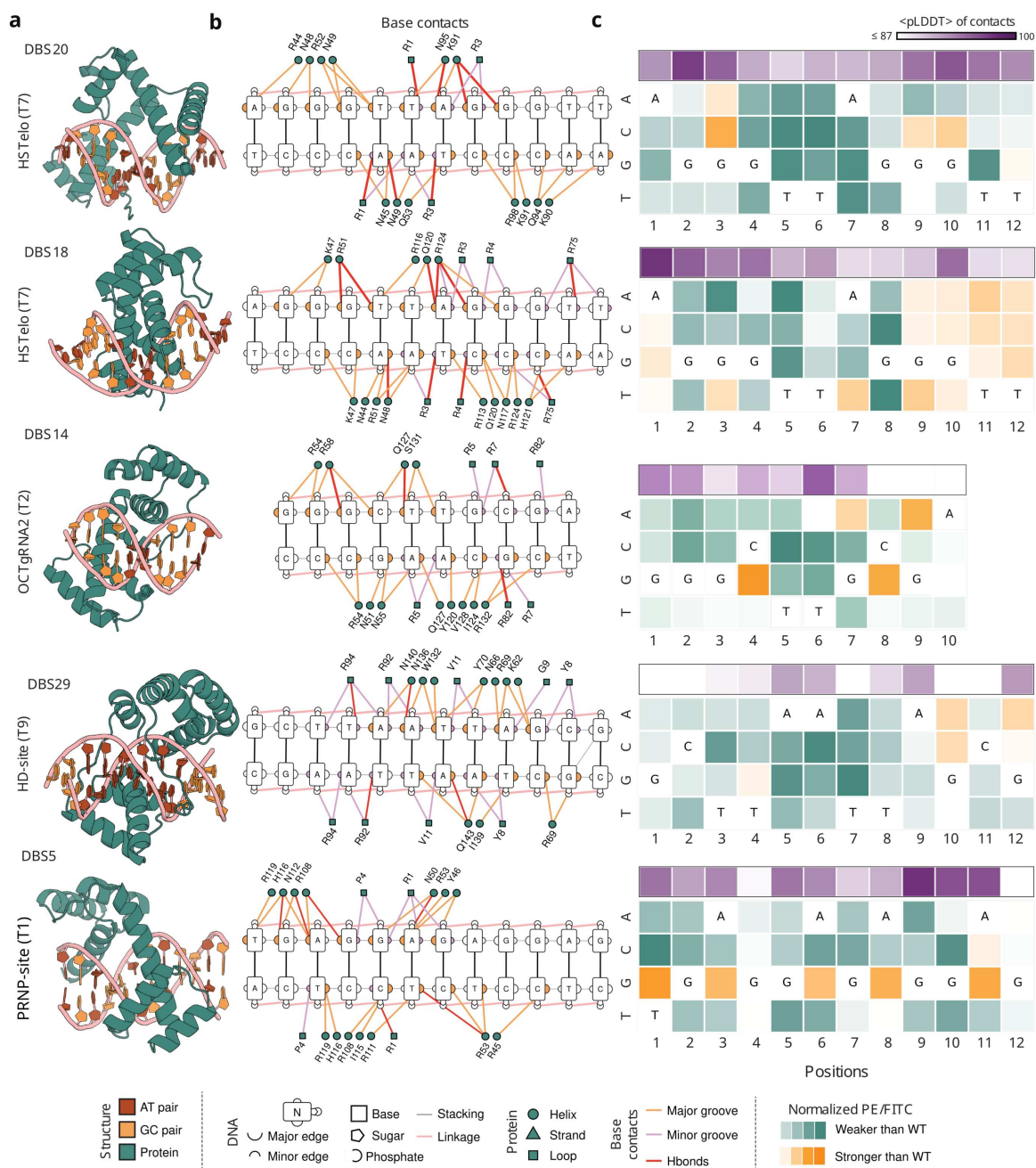


Figure 26

Figure 26 — Detailed base contacts of designs evaluated by variant competition assay. a, Design models (AF3 predictions) of selected specificity block designs characterized for fine-grained specificity using yeast surface display-based competition assay, assessing binding towards all possible single nucleotide variants of corresponding DNA target. b, Schematic representation of DNA base contacts (generated using DNAproDB [73] and modified stylistically) showing all protein residue contacts (with secondary structure annotation: helix, sheet, loop) to the major groove (orange) and minor groove (purple) side of the DNA bases. All Hydrogen bonds to the bases are highlighted in red. DNA backbone contacts are not shown. c, Results of competition assay (normalized PE/FITC) for each design (orange-green heatmap) presented in context of average AF3 pLDDT of contacting protein residues (within 4 Å) per base-pair position (purple heatmap). Green indicates a mutant competitor target was weaker than the intended (WT) base at that position, while orange indicates higher binding for the mutant competitor. Reproduced from Fig. S15 of Sehgal et al. [87].

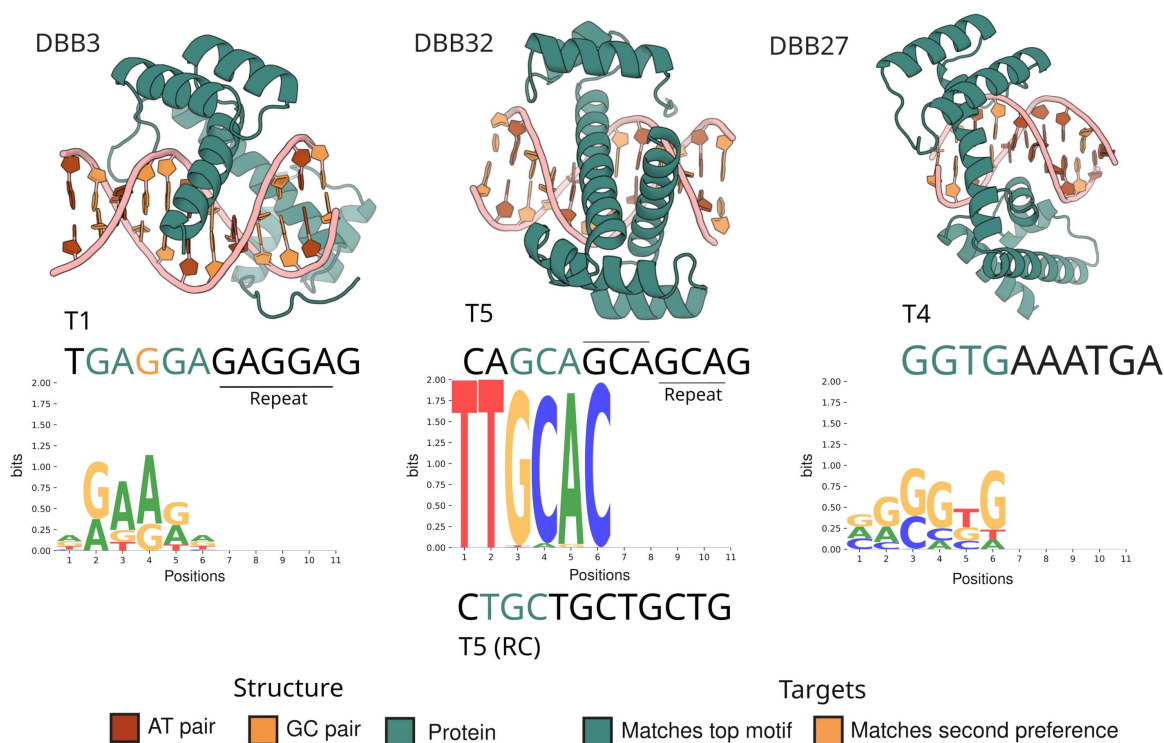


Figure 27 — HT-SELEX characterization of binder block designs. (Top of each row) Design models of binder block designs (DBBs) quantified via HT-SELEX [51, 81] with information content-rich specificity profiles replicated across experiments. (Bottom of each row) PWMs derived via HT-SELEX in context of the intended target sequences for each design. Green represents a match with the intended target. Orange represents intended bases where they appear as the second preference in the HT-SELEX derived PWM. The PWMs were derived from the count matrices with an uniform background assumption. Reproduced from Fig. S14 of Sehgal et al. [87].

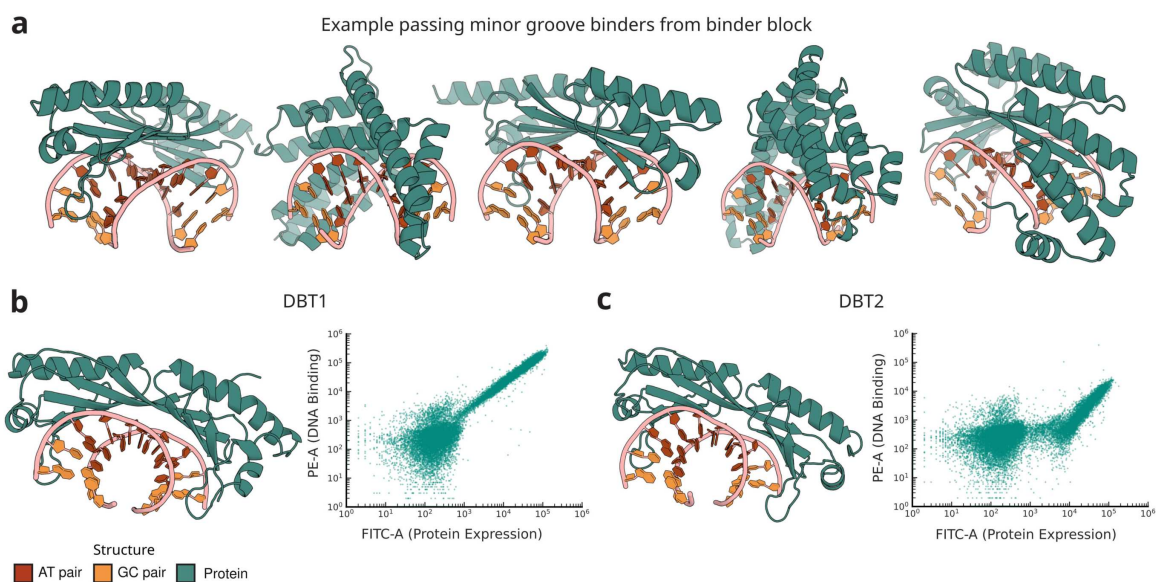


Figure 28 — Proof-of-principle design of minor groove DNA binding proteins. a, Designs derived from a slightly modified design pipeline targeting a bent TBP-site target. These designs are passing designs output from the binder block of the pipeline. b, DBT1, a design (model at left) that was found to be a binder via yeast surface display assay (right). This design has a best case 38% sequence similarity to PDB. c, DBT2, a design (model at left) that was found to be a binder via yeast surface display assay (right). This design has a best case 41% sequence similarity to PDB. Reproduced from Fig. S19 of Sehgal et al. [87].

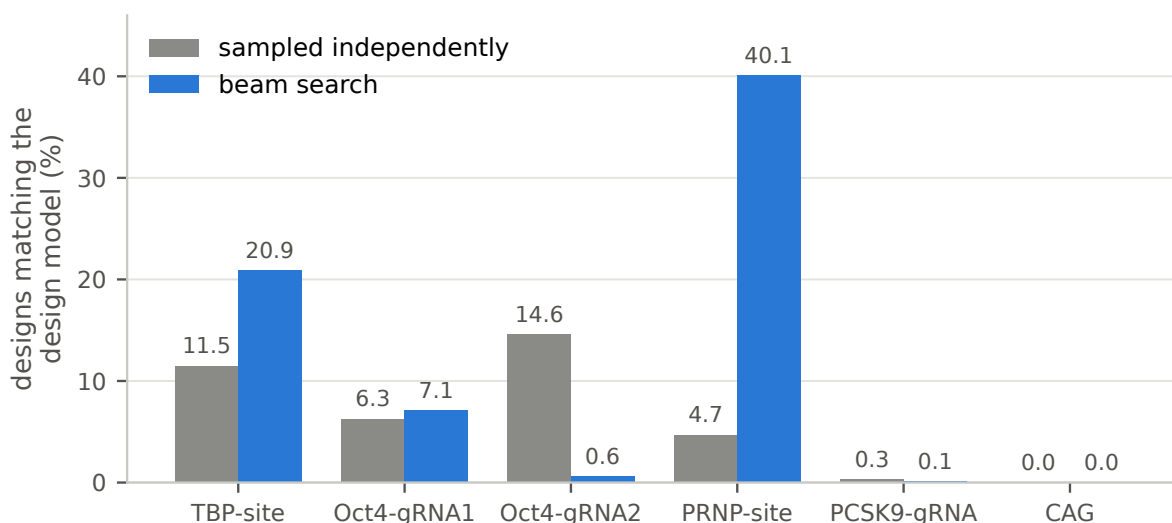


Figure 29 — On-target inference-time search across the six core DNA targets. Fraction of designs whose first-round AlphaFold3 prediction falls within 5 Å DNA-aligned protein C_α-RMSD of the design model, with terminal loops trimmed, for the partially latent flow model sampled independently (grey) and for the same checkpoint under beam search with the on-target interface minPAE as the reward (blue). Diffused protein length was sampled from 50 to 80 residues. Targets are the six core targets of Table 1. The comparison is between one training checkpoint with and without search, taken from a larger sweep that also covered two sequence-design routes and a hydrogen-bond-conditioning ablation. This work is ongoing and the values are preliminary.

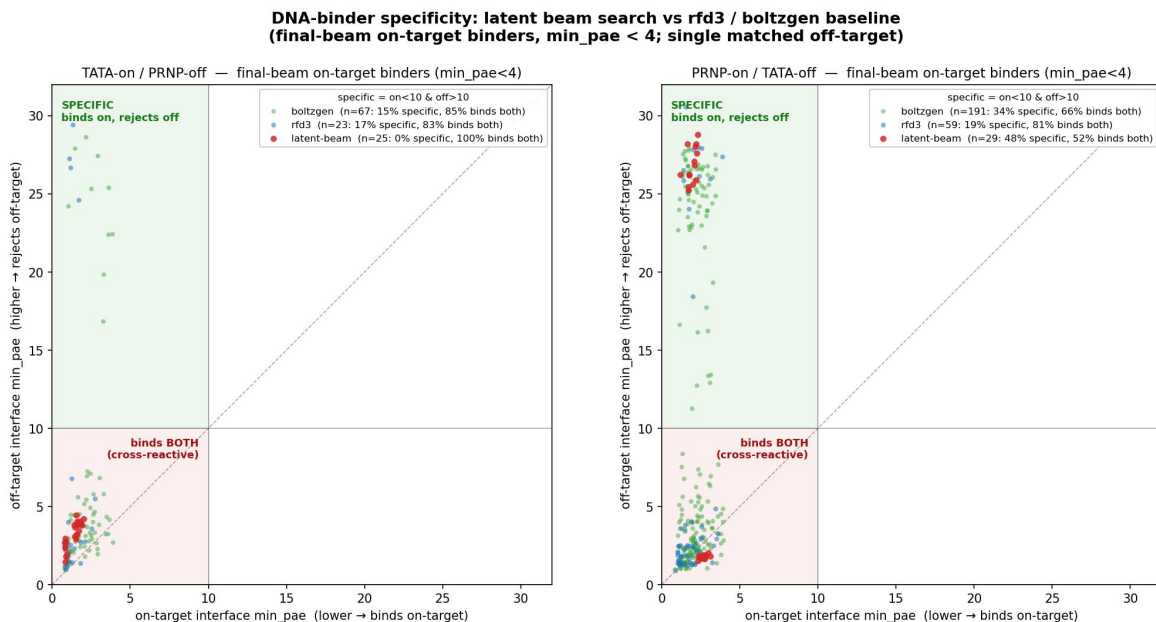


Figure 30 — Inference-time search under a combined on- and off-target reward. Interface minPAE of designs that bind their on-target *in silico* (on-target minPAE < 4), scored against the on-target sequence (x axis, lower means stronger predicted on-target binding) and against one matched off-target sequence (y axis, higher means stronger predicted off-target rejection), for BoltzGen (green), Rfdiffusion3 (blue) and the partially latent flow model under beam search with the combined reward (red). Left, designs against the TBP-site scored against the PRNP-site. Right, designs against the PRNP-site scored against the TBP-site. A design counts as specific when on-target minPAE < 10 and off-target minPAE > 10 (green region) and cross-reactive when both fall below 10 (red region). These are *in silico* rates defined by a pair of hard minPAE thresholds rather than measured specificities, and the series differ in size, from n = 23 to n = 191. This work is ongoing and the values are preliminary.

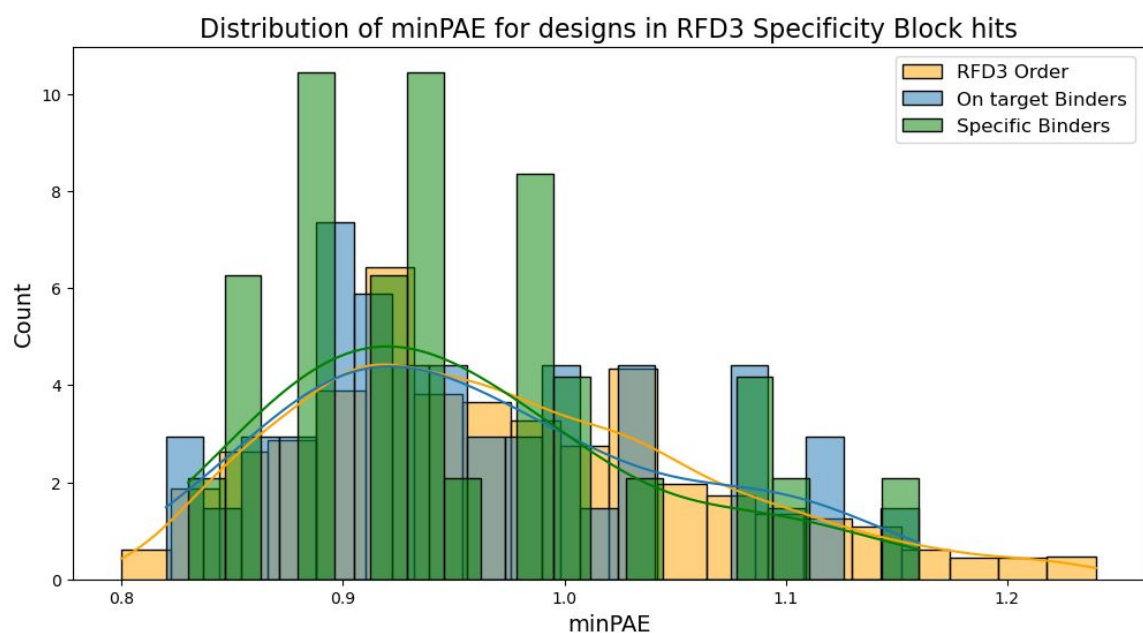


Figure 31 — Distribution of interface minPAE among ordered designs, on-target binders and specific binders. AlphaFold3 interface minPAE for the specificity block designs ordered in the RFDiffusion3 campaign (orange), for the subset that bound their on-target in the yeast display assay (blue), and for the subset that bound specifically (green), each histogram carrying a kernel density estimate. Every design shown had already passed the pipeline’s minPAE filter.

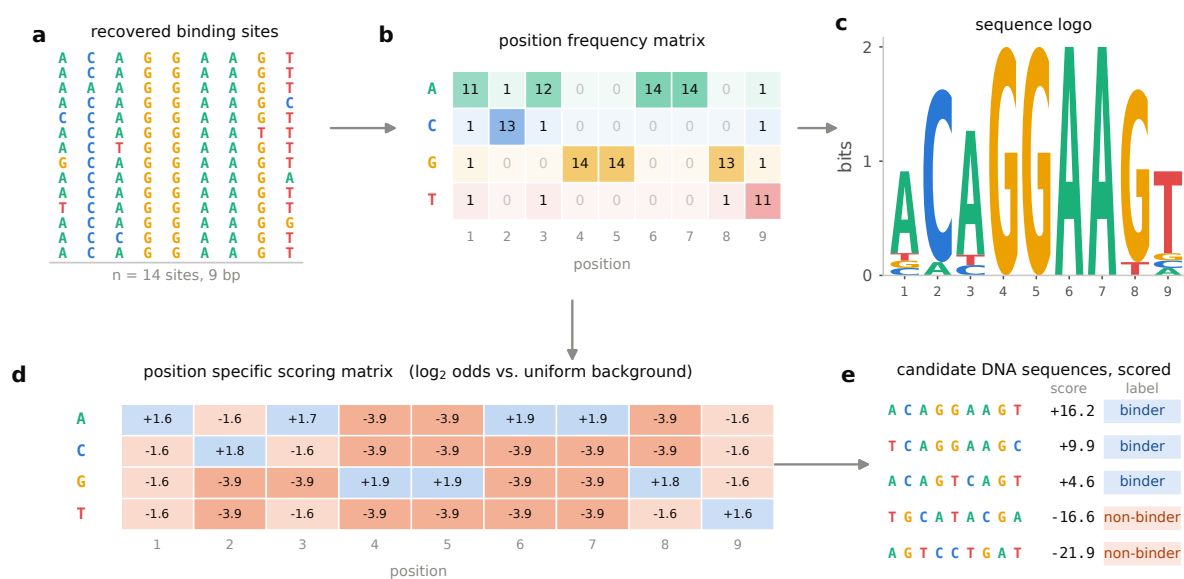


Figure 32 — How a position weight matrix becomes a training label. a, Example binding site count data recovered from an assay, aligned. b, Resulting position frequency matrix. c, The position frequency matrix expressed as a sequence logo. d, Scoring matrix, a log-odds against a uniform background, computed with a pseudocount of 0.25 per base. e, Candidate DNA sequences scored against that matrix. The score is used as the training label.

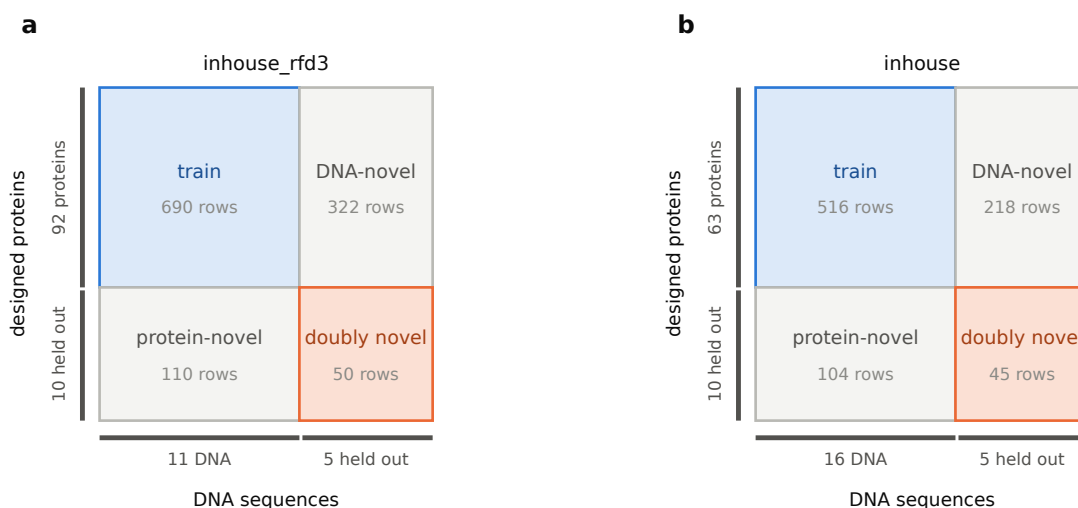


Figure 33 — The two-dimensional block split of the two in-house panels. Each panel measures a set of designed proteins against a set of DNA targets, and that matrix is cut into four quadrants by holding out ten proteins and five DNA sequences. The model is trained on the one quadrant whose protein and whose DNA both fall on the training side. a, *inhouse_rfd3*, the orthogonality panel from the RFDiffusion3 campaign of Chapter 3. b, *inhouse*, the panel from the earlier campaigns of Chapter 2. The five held-out DNA sequences were selected by greedy farthest-point sampling on 3-mer composition among the targets measured against at least 40% of the proteins in the panel, excluding any target that also appears in *dna_pwm*, *mitomi* or *rogers200*; the ten held-out proteins were selected among those measured against at least four of the five held-out targets, taking the ten with the largest range of measured intensity across them. The training quadrant and the doubly novel quadrant share no protein sequence and no DNA sequence in either panel.

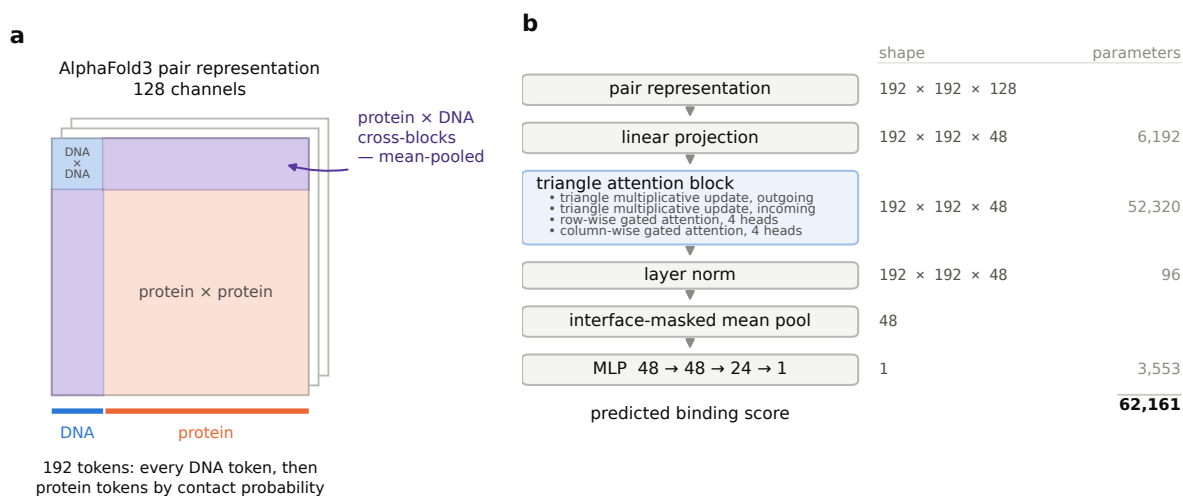


Figure 34 — The re-ranking model over AlphaFold3 pair embeddings. a, The crop of the AlphaFold3 pair representation the model takes as input, and the protein $\times$ DNA cross-blocks the network pools over. b, The stages of the network, with the output shape and the parameter count of each. Each of the four sublayers of the triangle attention block is output-gated with its gate initialized closed, so the block begins training near the identity.

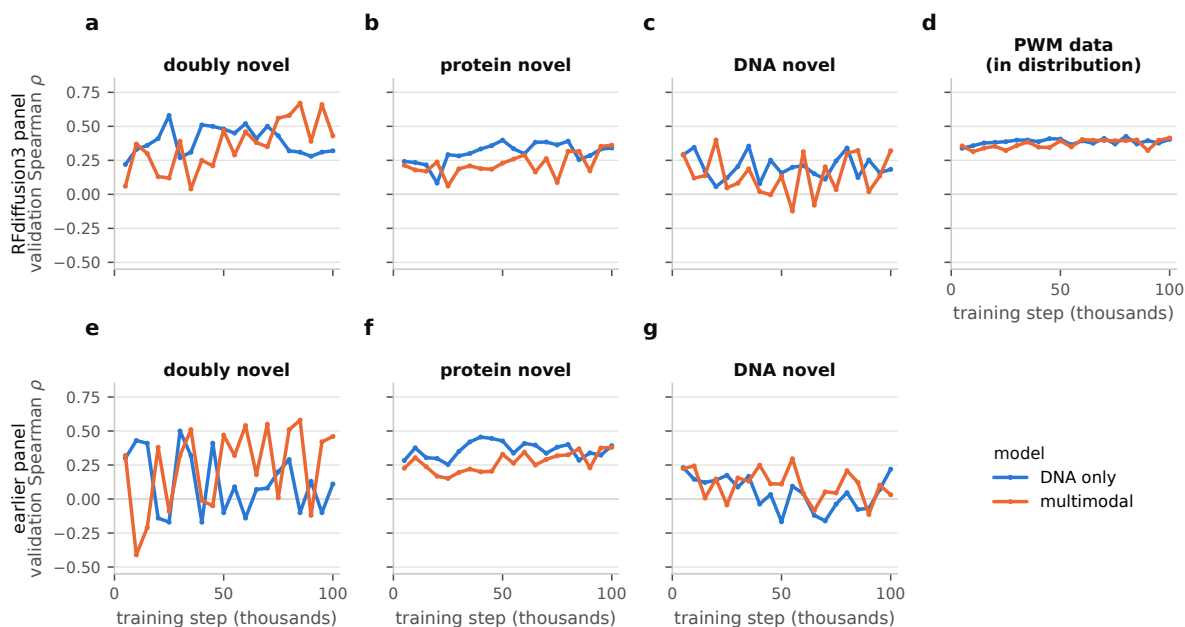


Figure 35 — Validation correlation over training for the DNA-only and multimodal models. Spearman correlation between the predicted and the measured values, recorded every 5,000 steps. a–c, The held-out quadrants of the RFdiffusion3 in-house panel (§4.2). d, The PWM-derived validation data, in distribution. e–g, The same quadrants of the earlier in-house panel. The doubly novel quadrants contain 50 rows over ten proteins in a and 45 rows over ten proteins in e, so a change of rank order in a handful of proteins moves those curves by the amount seen here. Both models are the `block5x10` checkpoints, and they differ only in the modality mix of their training batches.

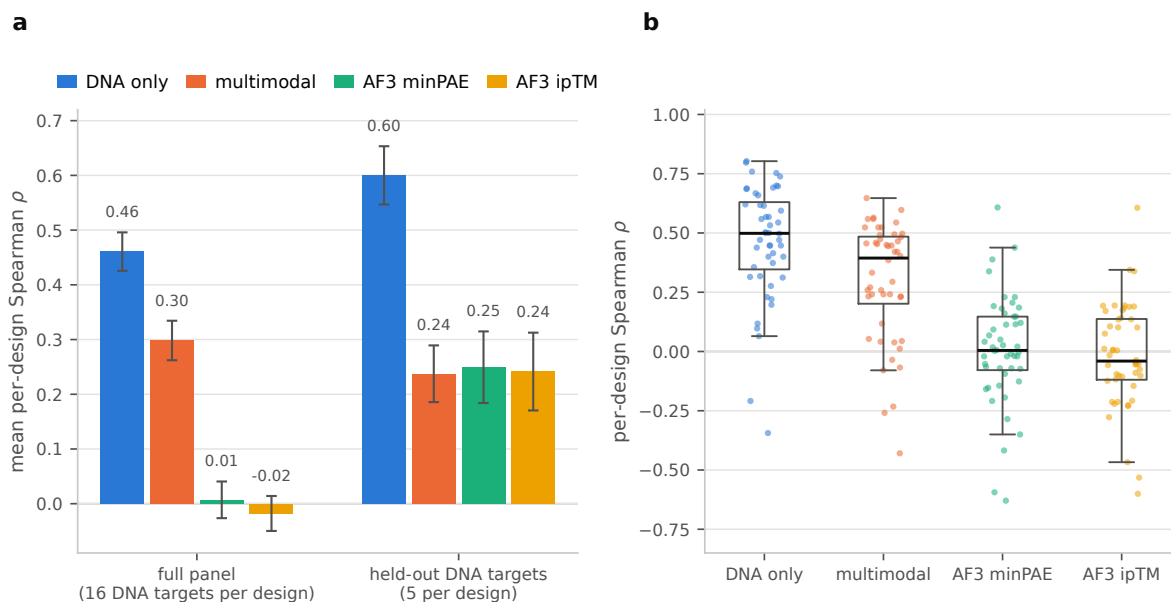


Figure 36 — Prospective evaluation on an orthogonality assay of 48 designs against 16 DNA targets. a, Spearman correlation between the predicted and the measured binding intensity, computed per design and averaged over the 48 designs. Error bars are the standard error of that mean. b, The distribution of those per-design correlations on the full panel, one point per design. On the held-out DNA targets the DNA-only model has a median per-design correlation of 0.70. All 48 designs postdate the training of both models, and each has at most 30% sequence identity to any protein in the training set.

Tables

Table 1 — The sixteen DNA design targets. Called from §3.2. Sequences and JASPAR accessions follow Table 1 of Sehgal et al. [87]. The core set was carried through both the binder block and the specificity block; the additional set was carried through the specificity block alone.

target	sequence	set	provenance
PRNP-site	TGAGGAGAGGAG	core	prion protein epigenetic editor site [77]
Oct4-gRNA2	GGGCTTGCGA	core	Oct4 guide RNA site
TBP-site	CGTATAAACG	core	JASPAR MA0108.3, TATA-binding proteins
Oct4-gRNA1	GGTGAAATGA	core	Oct4 guide RNA site
CAG	CAGCAGCAGCAG	core	Huntington’s disease trinucleotide repeat [6]
PCSK9-gRNA	ACCTTCTAGG	core	PCSK9 epigenetic editor site [92]
HSTelo	AGGGTTAGGGTT	additional	human telomeric repeat [75]
NFκB-site	GGGGATTCCCCC	additional	JASPAR MA0105.2, Rel homology region factors
HD-site	GCTTAATTAGCG	additional	JASPAR MA0094.2, homeodomain factors
Dux4-gRNA2	CAGGCCGCAGG	additional	DUX4 guide RNA site [41]
Ebox	ACCACGTGGT	additional	JASPAR MA0059.1, bHLH factors
P53-site	AGACATGTCT	additional	JASPAR MA0106.3, p53 domain factors
Tbox	AGGTGTGAAG	additional	JASPAR MA1919.2, T-box factors
FKH-site	GCGTAAACAA	additional	JASPAR MA0030.2, forkhead/winged helix factors
C4ZF-site	TGACCTTTGA	additional	JASPAR MA0017.1, nuclear receptors with C4 zinc fingers
Dux4-gRNA1	TCATCCAGCAG	additional	DUX4 guide RNA site [41]

Ebox is the one target for which no designs were ordered. HSTelo was filtered against the six core targets only, and every other additional target against all fifteen off-targets.

Table 2 — Design campaigns against DNA targets. Called from §4.1. Each campaign replaced the generative model, the filter, or both.

campaign	backbone generation	filter	bound any target	on-target and selective
Glasscock et al. [37]	RIFdock	Rosetta interface metrics	0.32%	0.027%
CA-RFdiffusion (§2.3)	alpha-carbon diffusion	RoseTTAFoldNA	4.4%	0.031%
RFdiffusion All-Atom (§2.3)	tip-atom guideposted diffusion	AlphaFold3	4.6%	0.027%
RFdiffusion3 (Chapter 3)	all-atom diffusion	AlphaFold3 minPAE and Δ minPAE	— [†]	roughly 3% of 96 designs per target

[†] The first three campaigns were characterised by pooled chip sorting, which gives one binding rate over all designs tested. The RFdiffusion3 campaign was characterised by individual eblock, which gives a rate per target: for the ten additional targets, binding rates ran from about 1% to about 19% and specificity rates from about 1% to about 14% (§3.4). The cell is left empty because the two quantities measure different things.

Table 3 — Curated protein-DNA datasets. Called from §4.2.

dataset	rows	protein sequences	DNA sequences	readout
dna_pwm	202,860	2,500	167,267	PWM-derived score
mitomi	5,600	28	1,423	inverse enrichment
rogers200	2,000	10	1,999	binary
inhouse	883	75	21	normalized intensity
inhouse_rfd3	1,172	102	16	normalized intensity

Table 4 — Composition of dna_pwm by source. Called from §4.2. Sits directly beneath Table 3 and describes the same data at a finer level. Kept as two tables, per your decision of 2026-08-05.

source	rows	protein sequences	DNA sequences
CisBP	92,221	1,479	84,422
JASPAR	69,373	702	59,498
Taipale HT-SELEX	17,423	215	17,308
ChIP-Atlas	17,296	63	16,158
UniProbe	6,547	41	6,475

Table 5 — Prospective evaluation on the 48-design orthogonality assay. Called from §4.4. Mean per-design Spearman correlation between predicted and measured intensity.

evaluation set	DNA-only	multimodal	AlphaFold3 minPAE	AlphaFold3 iPTM
full panel, 16 DNA targets per design	0.46	0.30	0.01	−0.02
held-out DNA targets, 5 per design	0.60	0.24	0.25	0.24