End-to-end single-stranded DNA sequence design with all-atom structure reconstruction

Yunda Si[1], Yaochen Xu[2], Luonan Chen[1,3,*]

[1]Key Laboratory of Systems Health Science of Zhejiang Province, School of Life Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China

[2]Center for Excellence in Molecular Cell Science , Chinese Academy of Sciences, Shanghai 200031, China

[3]School of Mathematical Sciences and School of AI, Shanghai Jiao Tong University, Shanghai 200240, China

*To whom correspondence should be addressed: lnchen@sjtu.edu.cn

Abstract

Designing biological sequences that fold into predefined conformations is a central challenge in bioengineering. Although deep learning has enabled significant advances in protein and RNA sequence design, progress in single-stranded DNA (ssDNA) design has been constrained by the limited availability of structural data. To address this challenge, we introduce InvDNA, a deep learning-based method that designs ssDNA sequences directly from backbone atomic coordinates. This end-to-end formulation avoids the loss of structural information during backbone-to-feature conversion and further accommodates flexible backbone representations, dynamic sequence masking, and structural reconstruction objectives. All these training strategies enhance the capacity of InvDNA to generalize across diverse ssDNA structural contexts while enabling additional functionalities, including generating diversity sequences for a given backbone, reconstructing base conformations from backbone and preserving functional sites. In benchmarks using experimentally determined ssDNA structures, InvDNA achieves over a twofold improvement in sequence recovery compared with existing ssDNA design approaches. Further computational validation using AlphaFold3 shows that 44.4% of InvDNA-designed sequences successfully fold into their predefined conformations. Notably, this success rate increases when backbone coordinates are perturbed to diversify the InvDNA-designed sequences. Collectively, these results establish InvDNA as a robust framework for rational ssDNA engineering.

Introduction

Single-stranded DNA (ssDNA) is an essential biomolecular class involved in key biological processes, including gene regulation[1], transcriptional control[2], and the catalysis of biochemical reactions[3,4]. As such, rationally engineered ssDNA can be used to detect and precisely modulate these processes, enabling wide-ranging applications in therapeutics[5,6], diagnostics[7], and biosensing[8,9]. Central to these capabilities is the ability to design sequences that reliably adopt a specified backbone conformation, providing a foundation for rational ssDNA engineering[10,11].

ViennaRNA[12,13] and NUPACK[14,15] are the most widely used computational frameworks currently available for ssDNA sequence design. Both approaches approximate the target backbone using secondary structure representations and employ empirical energy functions to identify sequences that achieve a minimum free energy state corresponding to the target secondary structure. However, using secondary structure as the primary design criterion, combined with the reliance on simplified energy functions, presents significant limitations. Secondary structures represent a simplified approximation of backbone geometry, so even if a designed sequence perfectly matches the target secondary structure, its three-dimensional structural fidelity cannot be guaranteed[16,17]. Furthermore, current energy functions approximate complex molecular interactions as linear combinations of a few empirical factors, such as the free energy of specific loop types, and derive their parameters from limited experimental

data[18–20]. This oversimplification fundamentally restricts both the accuracy and practical applicability of energy function-based approaches.

In protein and RNA sequence design, which are highly analogous to ssDNA sequence design, deep learning-based methods have largely replaced traditional energy function-based models, establishing a new design paradigm[21–25]. These deep learning-based methods encode backbone structures using detailed geometric descriptors, such as interatomic distances and torsion angles, and employ neural networks to identify sequences that fold into conformations consistent with these encoded geometric features. During training, the neural networks automatically capture high-dimensional correlations between geometric features and sequences. Although these strategies provide valuable guidance for developing ssDNA sequence design methods, their direct application is hindered by the limited availability of ssDNA structural data[26], relative to that of protein and RNA[22,27]. This data limitation poses a significant challenge to the generalization and robustness of deep learning model[28,29].

Flexible backbone representations and end-to-end deep learning offer a promising strategy for improving the generalization ability of deep learning-based sequence design models. Unlike fixed geometric features that encode backbones, which restrict the model to a single view of each structure-sequence pair, flexible backbone representations that randomly sample subsets of backbone atoms as encoded backbone enable analysis of each structure-sequence pair from a variety of perspectives. This flexibility has the potential to improve the expressiveness of learned representations and

enhance the predictive performance of the model. An end-to-end framework that operates directly on atomic coordinates naturally supports such flexible backbone representation and can also use the full backbone as input, thereby avoiding the loss of structural information during conversion from backbone to geometric features.

The goal of reconstructing all-atom structures from target backbones and designed sequences provides another promising strategy for improving the generalization ability of sequence design models by imposing explicit spatial constraints between designed sequences and their target backbones. Biomolecules with well-defined three-dimensional structures require sequence design methods capable of generating sequences whose all-atom conformations align with the target backbone while satisfying relevant physicochemical constraints. However, current deep learning approaches primarily focus on inferring amino acid or nucleotide identities from geometric features and do not explicitly enforce structural compatibility between designed sequences and target backbones. Several studies, including CarbonDesign[30], ABACUS-R[23], and LigandMPNN[31], incorporate side-chain conformation prediction as an auxiliary training objective and report measurable improvements. Despite these advances, such indirect modelling of spatial relationships does not fully capture local atomic interactions, such as those between backbone and base atoms, and therefore cannot completely prevent spatial clashes.

In this study, we introduce InvDNA, a deep learning-based method for ssDNA sequence design built on an end-to-end deep learning framework that inversely

transforms a predefined backbone conformation into an ssDNA sequence. Specifically, this framework takes backbone atoms and masked sequences as input and outputs designed ssDNA sequences along with all-atom structures. In addition to adopting training strategies including flexible backbone representation and all-atom structure reconstruction objectives, InvDNA incorporates dynamic sequence masking, where 0–20% of nucleotides in the input sequence are randomly retained during training. This retention simulates preserving functionally important nucleotides that may be constrained by their functional roles.

Benchmarking on experimental ssDNA structures shows that InvDNA outperforms traditional energy function-based ssDNA sequence design methods, achieving a twofold improvement in sequence recovery rate and a twofold to threefold increase in the proportion of designed sequences that fold into the target backbone conformation. These results demonstrate the effectiveness of the deep learning paradigm for ssDNA design. Moreover, the training strategies employed by InvDNA not only enhance sequence design accuracy but also enable additional capabilities, including generating diverse sequences for a given backbone, reconstructing base conformations from backbone and designing sequences that constrain specific nucleotide types at designated positions. Finally, given the limited availability of ssDNA structural data, we investigate how the number of ssDNA structures available for training influences the performance of InvDNA. Our study shows that the performance of InvDNA can be further improved by leveraging additional training data.

Results

Overview of InvDNA

The overall architecture of InvDNA is illustrated in Figure 1a. The framework takes masked sequences and backbone coordinates as direct inputs and uses a sequence representation that is iteratively updated through twelve structure blocks to predict both all-atom coordinates and nucleotide identities. For the sequence input, since each position in the ssDNA backbone contains multiple atom types, we model the sequence representation as a three-dimensional tensor that encodes atom-level information across inter-nucleotide, intra-nucleotide, and channel dimensions. The backbone coordinates are also represented using a three-dimensional tensor spanning inter-nucleotide, intra-nucleotide, and spatial dimensions. Coordinates of non-existent atoms (e.g., the “C7” atom in adenine) or atoms that are masked are replaced with the centroid of the existing atoms. By default, the backbone uses all backbone atoms for representation (see Figure 1b). When using flexible backbone representation, the atom types “P”, “C3'” and “C1'” are always retained as core atoms, while the remaining atoms are randomly masked to enhance structural diversity during training.

Each structure block receives both the sequence representation and backbone coordinates as inputs, while the initial sequence representation is derived from embeddings of the masked sequence. The block then updates the sequence representation using inter-nucleotide attention, intra-nucleotide attention, and transition modules sequentially. As shown in Figure 1c, the inter-nucleotide attention module,

inspired by the invariant point attention module proposed by AlphaFold2[32], updates the sequence representation by considering both the backbone coordinates and the relative nucleotide position within the inter-nucleotide dimension. The intra-nucleotide attention module has a similar architecture to inter-nucleotide attention module (see Figure S1) and updates the sequence representation based on intra-nucleotide interactions within the intra-nucleotide dimension. Finally, the transition module[33] updates the sequence representation through channel-wise transformations.

Performance of InvDNA on experimental ssDNA structures

We used recently released experimental ssDNA structures (referred to as the ssDNA dataset) to benchmark InvDNA against currently available ssDNA sequence design methods, including ViennaRNA (using DNA2004 free energy parameters) and NUPACK (using dna04.2 free energy parameters). The ssDNA dataset consists of 45 ssDNA structures released in the Protein Data Bank (PDB)[34] between June 2021 and June 2024, with each structure sharing less than 100% sequence identity with any sequence in InvDNA's training set (see Methods). Since ViennaRNA and NUPACK require secondary structures as input, we used DSSR[35] to extract secondary structures from the corresponding ssDNA three-dimensional structures.

We first evaluated the sequence recovery rates of sequences designed by the three methods on the ssDNA dataset. The sequence recovery rate is defined as the percentage of positions in the designed sequence that match the native sequence for a given

backbone; higher sequence recovery generally correlates with better structural fidelity to the target backbone. As shown in Figure 2a-b, for 40 out of the 45 ssDNA backbones, the sequences designed by InvDNA exhibit significantly higher or comparable sequence recovery relative to ViennaRNA and NUPACK, highlighting the superior performance of InvDNA's deep learning-based approach. Furthermore, Figure 2b shows that InvDNA achieves similar performance on backbones whose native sequences share 80-100% identity to InvDNA's training set, as well as on those with less than 80% identity, demonstrating the robustness of InvDNA to sequence divergence. In terms of mean sequence recovery rate, InvDNA achieves 57.9% on the ssDNA dataset, whereas ViennaRNA and NUPACK achieve only 28.2% and 24.5%, respectively.

We further assessed whether the designed sequences could fold into their intended backbone conformations using AlphaFold3[33], a state-of-the-art structure prediction model for biomolecular complexes. All AlphaFold3 runs employed the default settings. In addition to sequences generated by InvDNA, ViennaRNA, and NUPACK, we also predicted structures for the corresponding native sequences as a control. For each designed or native sequence, we calculated the root-mean-square deviation of C3' atoms (denoted as RMSD-C3') between the target backbone and the AlphaFold3-predicted structure, defining successful folding as an RMSD-C3' below 5 Å. As shown in Figure 2c, the success rate of sequences designed by InvDNA reaches 44.4%, approximately threefold higher than that of ViennaRNA (13.3%) and twofold higher than NUPACK (22.2%). A representative example is design sequence for the G-quadruplex (PDB:

8D79). As shown in Figure 2e (Figure S2 for more examples), the structure predicted from the InvDNA-designed sequence attains an RMSD-C3' of 1.2 Å, whereas the sequences designed by ViennaRNA and NUPACK yield RMSD-C3' values of 9.6 Å and 11.3 Å, respectively. These results collectively demonstrate the advantages of the deep learning-based design paradigm in ssDNA sequence design and highlight the utility of combining InvDNA with AlphaFold3. Notably, the success rate for native sequences is 40.0%, indicating that the evaluation is partly limited by AlphaFold3's predictive accuracy. The median RMSD-C3' values for each method on the ssDNA dataset are presented in Figure 2d: InvDNA attains a median RMSD-C3' of 6.18 Å, representing improvements of 3.21 Å and 3.06 Å over ViennaRNA and NUPACK, respectively. To mitigate the impact of AlphaFold3 prediction errors, we constructed a filtered dataset by excluding 27 ssDNAs for which AlphaFold3 failed to recover the correct structure even when provided with the native sequence (RMSD-C3' > 5 Å). On the remaining 18 ssDNAs, for which native sequences were accurately predicted, InvDNA exhibited even greater advantages, achieving a median RMSD-C3' of 3.30 Å, compared with 8.11 Å and 8.29 Å for ViennaRNA and NUPACK, respectively.

Given the importance of generating diverse sequences for a given backbone, we further assessed InvDNA's capacity to produce such diversity. InvDNA was trained with flexible backbone representations that perturbs the atom-type encoding of the input backbone, thereby naturally enabling variation in the resulting sequences. In addition, introducing Gaussian noise into the backbone atom coordinates offers an alternative

means of perturbation, an approach previously used in ProteinMPNN[22] to enhance design diversity. For each backbone in the ssDNA dataset, we generated 25 sequences using both the flexible-backbone strategy and the flexible-backbone–plus–coordinate-noise strategy. As shown in Figure 2f, the mean RMSD-C3' of the top -performing design (the sequence with the lowest RMSD-C3' among those generated) decreases sharply as the number of sampled sequences increases. Concurrently, Figure 2g shows that the mean number of unique sequences (defined as sequences those sharing less than 100% pairwise identity) rises steadily with deeper sampling. Notably, perturbing the backbone coordinates with Gaussian noise (standard deviation = 0.02 Å) substantially improves both structural accuracy and sequence diversity compared with using flexible-backbone representations alone. We also evaluated a larger noise magnitude (standard deviation = 0.04 Å); however, this did not yield further improvements (see Figure S3). Together, these results demonstrate that InvDNA effectively produces multiple structurally viable and sequence-diverse designs for a given backbone, thereby providing rich candidates for experimental validation and functional optimization.

Interpreting the InvDNA

We trained ablation models to evaluate the effectiveness of three key components involved in training InvDNA: flexible backbone representations, structure reconstruction objectives, and dynamic nucleotide masking. Specifically, the “No atom sampling” model was trained by replacing the flexible backbone representation with full backbone atoms as input. The “No structure loss” model was trained by removing

structure reconstruction related losses, including clash loss, bond loss and all-atom frame aligned point error (FAPE) loss (see Methods). The “No dynamic masking” model was trained by removing the strategy of random retaining 0-20% of nucleotides in input sequence during training, resulting in a fully masked sequence. After convergence, all models were evaluated on the ssDNA dataset, using fully masked sequence and complete backbone atoms as input during inference by default.

We first investigated the sequence recovery rate and RMSD-C3' of AlphaFold3-predicted structures for sequences generated by InvDNA and three ablation models, as shown in Figure 3a (Figure S4 for pairwise comparisons). Consistent with our hypothesis, removing flexible backbone representations or structure reconstruction objectives consistently decreased both sequence recovery rate and RMSD-C3', demonstrating that each component improves the generalization capability of sequence design model. Notably, the dynamic nucleotide masking strategy, which aims to simulates preserving functionally important nucleotides, also slightly enhances the model’s generalization ability. Specifically, InvDNA achieves median sequence recovery rate and RMSD-C3' of 56.3% and 6.18 Å. The “No atom sampling” model achieves median sequence recovery rate and RMSD-C3' of 51.9% and 8.29 Å, the "No structure loss" model achieves median sequence recovery rate and RMSD-C3' of 50.0% and 7.92 Å, and the "No dynamic masking" model achieves median sequence recovery rate and RMSD-C3' of 54.5% and 8.90 Å. The mean sequence recovery rate and RMSD-C3' values for each model are shown in Table S1.

To investigate the ability of InvDNA to preserve specified nucleotides as one of its training objectives, we analyzed the sequence recovery rate of sequences designed by InvDNA and the “No dynamic masking” model while systematically varying the proportion of unmasked nucleotides in input sequence (from 0% to 50%). As shown in Figure 3b, the sequence recovery rate of InvDNA increases nearly linearly with the percentage of unmasked nucleotides, reaching 99.2% when the complete native sequence is provided as input. In contrast, the "No dynamic masking" model that trained on fully masked input sequences, is unable to effectively utilize the provided sequence information, resulting in a slight decrease in sequence recovery rate as more nucleotides are unmasked. This result demonstrates that the dynamic nucleotide masking strategy is essential for the model to effectively incorporate partial sequence constraints.

We also investigated the ability of InvDNA to reconstruct physically realistic all-atom structure from the backbone, as another important training objective. The quality of the all-atom structures reconstructed during training on the validation dataset is shown in Figure 3c-e (Mean sequence cross-entropy loss and sequence recovery rate are shown in Figure S5). Specifically, Figure 3c shows the mean RMSD-C3' of reconstructed structures relative to the input backbone, while Figure 3d and 3e show the mean bond length deviation and the number of steric clashes, respectively. All three metrics improve substantially during training: the RMSD-C3' decreases from an initial value of 6.5 Å to 0.007 Å, bond length deviation decreases from 5.8 Å to 0.14 Å, and the number of steric clashes decreases from nearly 500 to nearly 175 per prediction.

This substantial improvement confirms that the structure reconstruction-related losses effectively guide the model in predicting physically plausible all-atom structures from the backbone.

Although InvDNA is not specifically designed to reconstruct base conformations from full backbones and native sequences, the convergence of structural reconstruction loss and the ability to retain core nucleotides suggests that InvDNA may possess the capability to accurately construct base conformations when both the full sequence and the backbone are provided as input. Figure 3f shows the quality of structures generated by InvDNA when provided with both the full sequence and the backbone on the ssDNA dataset. Unlike the previously used RMSD-C3' metric for structure evaluation, we employed the Local Distance Difference Test (LDDT)[36] to assess the accuracy of the predicted structure at the atomic level. LDDT evaluates the accuracy of atomic positions by considering atomic microenvironments. An LDDT value of 1 indicates that the atomic microenvironment in the predicted structure is nearly identical to that in the experimental structure. As shown in Figure 3e, the median LDDT for atoms on the bases reached nearly 0.9, indicating that most atoms on the bases were located at the correct position. For atoms on the backbone, the median LDDT further reaches 0.95, indicating extremely high fidelity of the backbone. Figure 4 shows structures generated by InvDNA with both full backbone and native sequence input. For ssDNA with different structural types or different functions, bases in most predicted structures have the correct orientation relative to native structure (Figure 4a) and even hydrogen bonds between

bases are correctly modelled (Figure 4b), further indicating that most atoms on the bases were located at the correct position. It should be noted that the structures generated by InvDNA still contain physical inconsistencies (Figure 4c for clash error and Figure 4d for bond length error), which can be eliminated by combining with molecular dynamics simulations. Notably, we trained a dedicated InvDNA version (denoted as "full sequence") to reconstruct base conformations by inputting complete sequences and backbones during training. As shown in Figure 3g, this model achieved a significant improvement in base conformation accuracy compared to the original InvDNA (Figure S6 shows the quality of structures generated by "full sequence" model at the atomic level).

Training data size analysis

Given the limited availability of ssDNA structural data, we also investigated how the amount of available ssDNA structures for training affects the performance of InvDNA. Specifically, we employed the same training methodology as InvDNA to train three additional ablation models, each using 12.5%, 25%, and 50% of the total clusters from the training data. The ssDNAs in the training dataset of InvDNA were clustered using CD-HIT[37] with sequence identity 80%, result into 255 clusters. After convergence, all models were evaluated on the ssDNA dataset, using fully masked sequence and complete backbone atoms as input during inference.

Figure 5a shows the median and mean sequence recovery rates of sequences generated by InvDNA and the three ablation models. The performance of sequence design models improves as the availability of ssDNA structural data increases, demonstrating that InvDNA's performance is constrained by the amount of available ssDNA structures. Pairwise comparisons between the ablation models and InvDNA, shown in Figure 5d-f, indicate that the increase is systematic. We further evaluated these models by assessing the quality of AlphaFold3-predicted structures using generated sequences. As shown in Figure 5b-c, the mean LDDT (calculated using C3' atoms) and mean RMSD-C3' for the sequence design models also increase with more available ssDNA structural data (pairwise comparisons between ablation models and InvDNA are shown in Figure S7). The median and mean sequence recovery rate, LDDT and RMSD-C3' values for each model are shown in Table S2. Notably, when trained with just 12.5% of the full training clusters, InvDNA still significantly outperforms traditional methods in sequence recovery rate, demonstrating the effectiveness of deep learning approaches for ssDNA sequence design.

Discussion

In this work, we propose InvDNA, a deep learning-based method for ssDNA sequence design. InvDNA employs an end-to-end deep learning framework that takes backbone atoms and masked sequences as input and outputs designed ssDNA sequences along with all-atom structures. Based on this framework, we also introduce several training

strategies, including flexible backbone representations, structure reconstruction objectives, and dynamic nucleotide masking, to enhance sequence generation and enable additional functionalities.

Experiments using experimentally determined ssDNA structures demonstrate that InvDNA consistently outperforms existing ssDNA sequence design methods across multiple scenarios. Ablation experiments demonstrate that the training strategies utilized by InvDNA further enhance its performance. Additionally, InvDNA can generate diverse and structurally viable sequences for a given backbone, accurately reconstruct base conformations from the backbone and preserve specified nucleotides in the input sequence. We also investigated the impact of the number of available ssDNA structures on the performance of InvDNA. Our results show that even when ssDNA structures are extremely limited, InvDNA outperforms traditional ssDNA sequence methods in sequence recovery rate, underscoring the effectiveness of deep learning approaches for ssDNA sequence design.

Reliable operation of InvDNA still depends on a complete backbone, and minor physical inconsistencies remain in the reconstructed all-atom structures. Addressing these challenges presents an opportunity to further improve the generalization and applicability of InvDNA. Additionally, our work is limited to computational analysis of the designed sequences, wet-lab validation is crucial for a comprehensive assessment of InvDNA's reliability and may offer new insights and opportunities for improvement. Besides, Although InvDNA is primarily designed for ssDNA sequence design, its

ssDNA-independent framework and training strategies allow it to be easily extended to protein, RNA, and other biological sequence design tasks.

Methods

Training and validation datasets

All DNA chains deposited in the Protein Data Bank (PDB) prior to 30 December 2020 were retrieved to construct the training dataset for InvDNA. Given the scarcity of single DNA chain with stable conformations (i.e., those not in close spatial proximity (within 10.0 Å) to other biomolecules), as most DNA chains in PDB exist in complex form, we collected DNA chains with at least locally stable conformations to construct the training dataset (similar to NuFold[38]). Specifically, DNA chains were retained if they met the following criteria: (1) resolution better than 9 Å if a reported resolution was available; (2) no non-standard residues; (3) at least four resolved residues; and (4) at least four non-local contacts. A non-local contact is defined as minimal interatomic distance (heavy atoms) between nucleotide i and nucleotide j of less than 5 Å, with |j-i| greater than 3. All the collected structures with identical sequences were clustered, yielding 480 unique DNA sequences. This collection, referred to as the "experimental dataset", served as the training dataset for InvDNA.

DNA chains deposited in the PDB between 30 December 2020 and 30 December 2021 were used to construct the validation dataset. After applying the same filtering

criteria as for the experimental dataset, the remaining chains were clustered with those from the experimental dataset using CD-HIT-EST-2D (-i2 validation.fasta -i train.fasta -c 1 -M 0 -T 0 -g 1 -n 2 -d 60 -G 1 -r 0 -l 1) at 100% sequence identity to eliminate redundancy. From clusters composed exclusively of candidate validation DNA chains, reference monomers were retained. This process resulted in a validation dataset comprising 27 unique DNA sequences. It is worth noting that, for DNA sequences in both the training and validation dataset that without structure in monomeric form, nucleotides lacking non-local contacts were masked during model training.

Test dataset

DNA chains deposited in the PDB between 30 December 2021 and 17 June 2024 were used to construct the test dataset. After applying nearly the same filtering criteria as for the experimental dataset, with the exception of a stricter requirement that at least 80% of the base sites in the DNA have at least one non-local contact, the remaining chains were clustered with those from both the experimental and the validation dataset using CD-HIT-EST-2D (-i2 test.fasta -i train+valid.fasta -c 1 -M 0 -T 0 -g 1 -n 2 -d 60 -G 1 -r 0 -l 1) at 100% sequence identity to eliminate redundancy. From clusters composed exclusively of candidate test DNA chains, reference monomers were retained. This process resulted in a validation dataset comprising 45 unique DNA sequences.

Training protocol

We trained InvDNA and ablation models on an H100 GPU with automatic mixed precision, exponential moving average, and a batch size of 1. The AdamW optimizer, with a weight decay of 0.01 and a learning rate of 0.001, was used for model optimization. The model was trained for up to 100 epochs and the weights that exhibited the best performance on the validation dataset, based on sequence cross-entropy loss, were saved as the final model. During training, ssDNAs were randomly sampled from the experimental dataset. To improve training stability, we applied gradient clipping by global norm with a threshold of 0.1 on the parameters for each training example.

The loss function for InvDNA consists of sequence-related cross-entropy loss and structure-related losses, including clash loss, bond loss and FAPE loss[32]. The cross-entropy loss aimed to restore native sequences: $\mathcal{L}_{seq} = \frac{1}{N}\sum_{n=1}^{N}\sum_{b=1}^{4} y_n^b log\, p_i^b$, where y_n denote one-hot encodings of nucleotide, and the bin probabilities p_i^b are obtained by linearly project the sequence representations into 4 bins and normalization using a softmax function. The clash loss penalizes inter-atom distances shorter than a defined tolerance: $\mathcal{L}_{clash} = \frac{1}{N}\sum_{i=1}^{A}\sum_{J=1}^{A} max(min(y_{ij}, 3.0) * 0.6 - d_{ij}, 0)$, where N is the number of nucleotides in the sequence, A is the number of heavy atoms in the structure. and y_{ij} and d_{ij} denote the interatomic distances in the experimental and predicted structure, respectively. The tolerance of atom overlap is set to 0.6. The bond loss enforces bond lengths within each nucleotide to remain within a reasonable length:

$\mathcal{L}_{bond} = \frac{1}{B}\sum_{i=1}^{B}|y_i - d_i|$, where B is the number of bonds in the structure. and y_i and d_i denote the bond lengths in the experimental and predicted structure, respectively. Considering that the convergence rate of structural losses is much slower than that of sequence loss, we reduced the weight of sequence loss. The corresponding weights for each loss component are set as 0.01 for cross-entropy loss, 0.05 for the clash loss, 0.05 for the bond loss, and 1 for the FAPE loss.

Evaluation Metrics

The quality of the designed sequence and the AlphaFold3-predicted or InvDNA-predicted structures using designed and native sequence was assessed using the sequence recovery rate, LDDT, and RMSD metrics.

The sequence recovery rate was calculated according to:

$$SR = \frac{1}{N}\sum_{n=1}^{N}(P_n == S_n)$$

Where N is the number of positions in backbone, P denote the designed sequence and S denote the native sequence. The LDDT and RMSD between predicted structures and predefined backbones was evaluated by the OpenStructure[39] package.

Benchmark Methods

ViennaRNA was collected from the latest version, 2.7.0, available on GitHub: https://github.com/ViennaRNA/ViennaRNA. NUPACK was collected from the latest version 4.0.2.0, available at https://www.nupack.org/download/software. AlphaFold3 was collected from GitHub: https://github.com/google-deepmind/alphafold3.

Acknowledgements

We thank the developers and contributors of open-source tools and datasets including PDB, PyTorch and others for greatly accelerating our research. This work was supported by National Key R&D Program of China (2022YFA1004800, 2025YFF1207900), Natural Science Foundation of China (T2350003, T2341007, 12131020, 42450084, 42450135, 12326614, and 12426310), Zhejiang Province Vanguard Goose-Leading Initiative (2025C01114), Science and Technology Commission of Shanghai Municipality (23JS1401300), and Hangzhou Institute for advanced study of UCAS (2024HIAS-P004), and JST Moonshot R&D (JPMJMS2021).

Author contributions

L.C. and Y.S. conceived the study. Y. S. performed the experiments and analyzed the results. L.C. and Y.S. wrote the manuscript. Y.X. developed the web server for InvDNA. L.C. supervised the research and provided funding.

Competing interests

The authors declare no competing interests.

Data and Code availability

The source code and pretrained weights of InvDNA will be released at https://github.com/yunda-si/InvDNA and https://biostructure.sibcb.ac.cn upon acceptance of the article.

References

1. Tian, Z. *et al.* Circular single-stranded DNA as a programmable vector for gene regulation in cell-free protein expression systems. *Nat Commun* **15**, 4635 (2024).
2. Gozzi, K., Salinas, R., Nguyen, V. D., Laub, M. T. & Schumacher, M. A. ssDNA is an allosteric regulator of the C. crescentus SOS-independent DNA damage response transcription activator, DriD. *Genes Dev* **36**, 618–633 (2022).
3. Hollenstein, M. DNA Catalysis: The Chemical Repertoire of DNAzymes. *Molecules* **20**, 20777–20804 (2015).
4. Chinnapen, D. J.-F. & Sen, D. A deoxyribozyme that harnesses light to repair thymine dimers in DNA. *Proceedings of the National Academy of Sciences* **101**, 65–69 (2004).
5. Scharner, J. & Aznarez, I. Clinical Applications of Single-Stranded Oligonucleotides: Current Landscape of Approved and In-Development Therapeutics. *Molecular Therapy* **29**, 540–554 (2021).
6. Nimjee, S. M., White, R. R., Becker, R. C. & Sullenger, B. A. Aptamers as Therapeutics. *Annu Rev Pharmacol Toxicol* **57**, 61–79 (2017).
7. Malicki, S. *et al.* Development of selective ssDNA micro-probe for PD1 detection as a novel strategy for cancer imaging. *Sci Rep* **14**, 28652 (2024).
8. Raveendran, M., Lee, A. J., Sharma, R., Wälti, C. & Actis, P. Rational design of DNA nanostructures for single molecule biosensing. *Nat Commun* **11**, 4384 (2020).
9. Sequeira-Antunes, B. & Ferreira, H. A. Nucleic Acid Aptamer-Based Biosensors: A Review. *Biomedicines* **11**, 3201 (2023).
10. Feldkamp, U. & Niemeyer, C. M. Rational design of DNA nanoarchitectures. *Angew Chem Int Ed Engl* **45**, 1856–1876 (2006).

11. Ashrafuzzaman, M., Tseng, C.-Y., Kapty, J., Mercer, J. R. & Tuszynski, J. A. A Computationally Designed DNA Aptamer Template with Specific Binding to Phosphatidylserine. *Nucleic Acid Ther* **23**, 418–426 (2013).

12. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms for Molecular Biology* **6**, 26 (2011).

13. Hofacker, I. L. *et al.* Fast folding and comparison of RNA secondary structures. *Monatsh Chem* **125**, 167–188 (1994).

14. Zadeh, J. N. *et al.* NUPACK: Analysis and design of nucleic acid systems. *J Comput Chem* **32**, 170–173 (2011).

15. Zadeh, J. N., Wolfe, B. R. & Pierce, N. A. Nucleic acid sequence design via efficient ensemble defect optimization. *Journal of Computational Chemistry* **32**, 439–452 (2011).

16. Vicens, Q. & Kieft, J. S. Thoughts on how to think (and talk) about RNA structure. *Proceedings of the National Academy of Sciences* **119**, e2112677119 (2022).

17. Abraham, M., Dror, O., Nussinov, R. & Wolfson, H. J. Analysis and classification of RNA tertiary structures. *RNA* **14**, 2274–2289 (2008).

18. Improved free-energy parameters for predictions of RNA duplex stability. https://www.pnas.org/doi/epdf/10.1073/pnas.83.24.9373 doi:10.1073/pnas.83.24.9373.

19. Jaeger, J. A., Turner, D. H. & Zuker, M. Improved predictions of secondary structures for RNA. *Proc Natl Acad Sci U S A* **86**, 7706–7710 (1989).

20. Xia, T. *et al.* Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719–14735 (1998).

21. Tan, C. *et al.* R3Design: deep tertiary structure-based RNA sequence design and beyond. *Brief Bioinform* **26**, bbae682 (2025).

22. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).

23. Liu, Y. *et al.* Rotamer-free protein sequence design based on deep learning and self-consistency. *Nat Comput Sci* **2**, 451–462 (2022).

24. Joshi, C. K. *et al.* gRNAde: Geometric Deep Learning for 3D RNA inverse design. *bioRxiv* 2024.03.31.587283 (2025) doi:10.1101/2024.03.31.587283.

25. Anand, N. *et al.* Protein sequence design with a learned potential. *Nat Commun* **13**, 746 (2022).

26. Zhang, Y., Xiong, Y. & Xiao, Y. 3dDNA: A Computational Method of Building DNA 3D Structures. *Molecules* **27**, 5936 (2022).

27. Wang, J., Wang, J., Huang, Y. & Xiao, Y. 3dRNA v2.0: An Updated Web Server for RNA 3D Structure Prediction. *International Journal of Molecular Sciences* **20**, 4116 (2019).

28. Hong, X., Zhan, J. & Zhou, Y. On the completeness of existing RNA fragment structures. 2024.05.06.592843 Preprint at https://doi.org/10.1101/2024.05.06.592843 (2024).

29. Ahdritz, G. *et al.* OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nat Methods* **21**, 1514–1524 (2024).

30. Ren, M., Yu, C., Bu, D. & Zhang, H. Accurate and robust protein sequence design with CarbonDesign. *Nat Mach Intell* **6**, 536–547 (2024).

31. Dauparas, J. *et al.* Atomic context-conditioned protein sequence design using LigandMPNN. *Nat Methods* **22**, 717–723 (2025).

32. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

33. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

34. Berman, H. M. *et al*. The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).

35. Lu, X.-J., Bussemaker, H. J. & Olson, W. K. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res* **43**, e142 (2015).

36. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).

37. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**, 282–283 (2001).

38. Kagaya, Y. *et al*. NuFold: end-to-end approach for RNA tertiary structure prediction with flexible nucleobase center representation. *Nat Commun* **16**, 881 (2025).

39. Biasini, M. *et al*. OpenStructure: an integrated software framework for computational structural biology. *Acta Cryst D* **69**, 701–709 (2013).

Figures

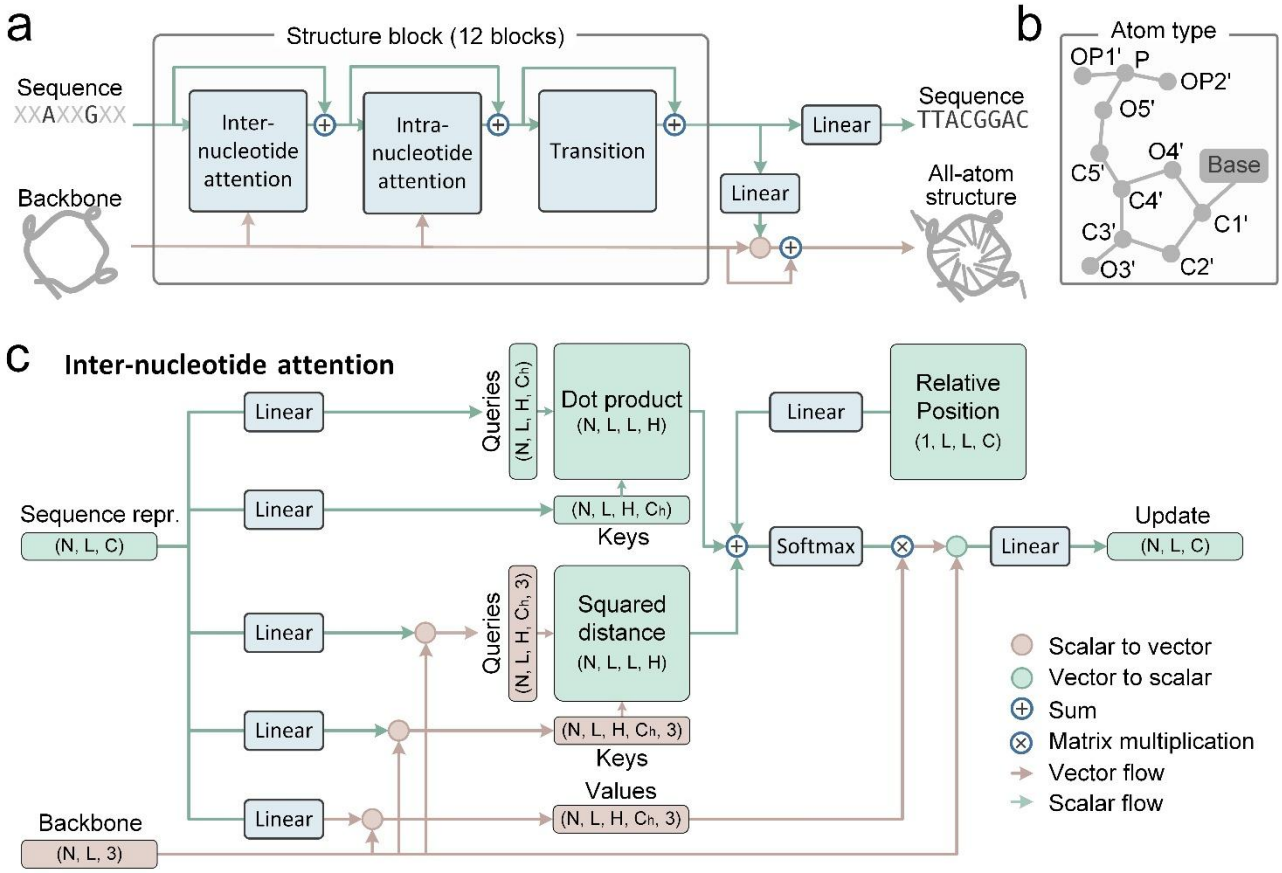

Figure 1: Overview of the InvDNA method. a, The end-to-end framework of InvDNA consists of 12 structure blocks, each block includes inter-nucleotide attention (Figure 1c), intra-nucleotide attention (Figure S1), and transition module[33]. Linear refer to linear layer in neural network. The input to the framework includes backbone coordinates, and, optionally, several core nucleotides. The output is a designed ssDNA sequence along with its corresponding all-atom structural model. **b,** Visualization of atom types used for backbone representation in InvDNA. **c,** Architecture of the inter-nucleotide attention module. The module takes sequence representation and backbone coordinates as input and outputs an updated sequence representation by learning inter-nucleotide interactions. Here, N denotes the number of atoms, L the sequence length, C the number of channels (default 384), and H the number of attention heads (default 12).

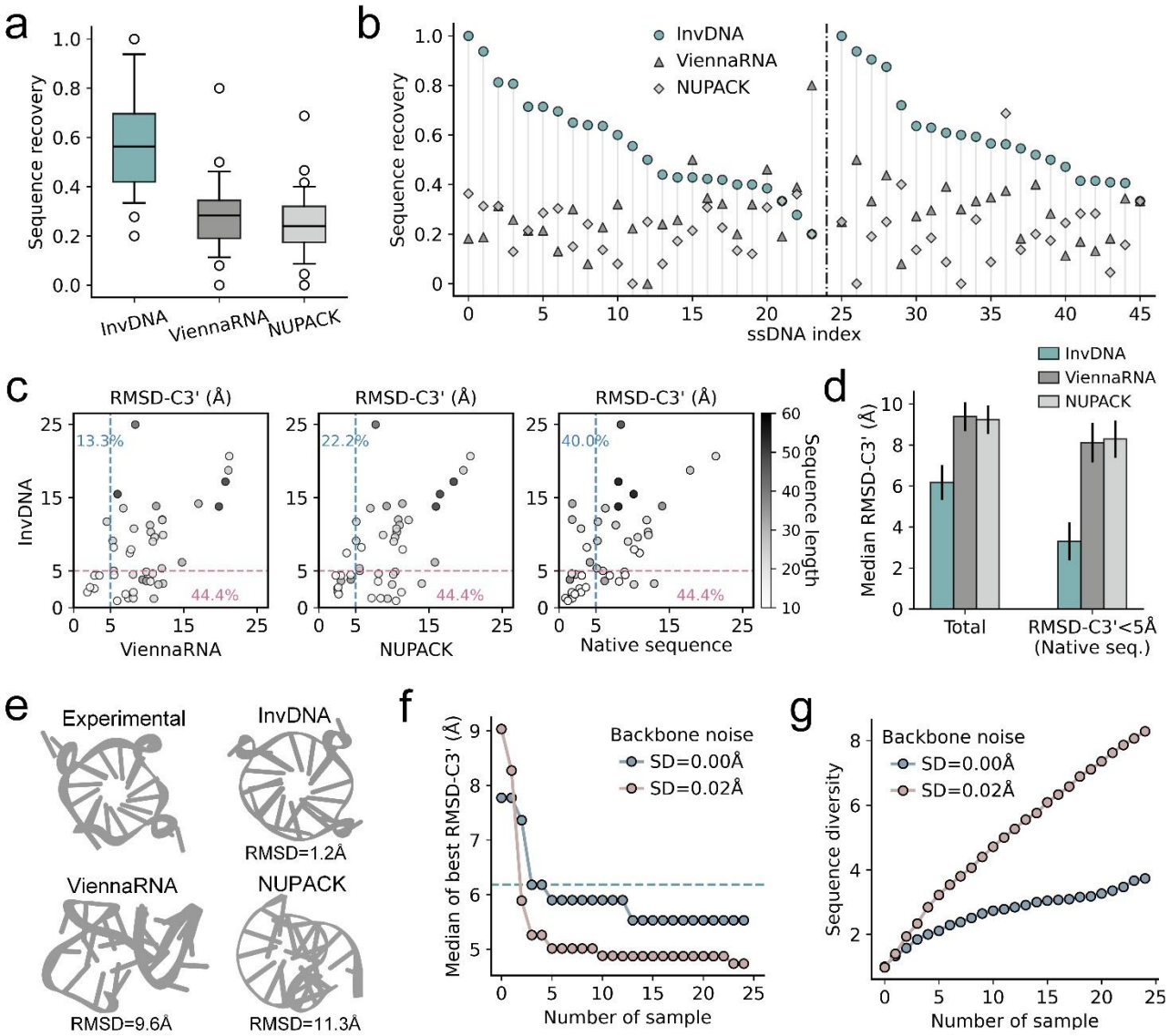

Figure 2: Performance of InvDNA on the ssDNA test set. a, Distribution of sequence recovery rates for sequences designed by InvDNA, ViennaRNA, and NUPACK on the ssDNA dataset. Box plots denote the 25–75% confidence intervals (box limits) and the median (center line). **b,** Sequence recovery rates for sequences designed by InvDNA, ViennaRNA, and NUPACK for each ssDNA backbone in the ssDNA dataset. ssDNAs to the left of the black dashed line share 80% to 100% with those in the InvDNA training set, whereas those to the right exhibit <80% similarity. **c,** Pairwise comparisons between structures predicted from InvDNA-designed sequences (y-axis) and those predicted from ViennaRNA-designed, NUPACK-designed, or native sequences (x-axis) on the ssDNA dataset. Structures are predicted using AlphaFold3, and similarity to the target backbone is evaluated

using RMSD-C3'. **d,** Mean RMSD-C3' of structures predicted using sequences generated by InvDNA, ViennaRNA, NUPACK on the full ssDNA dataset and on the subset for which AlphaFold3 produces accurate predictions (RMSD-C3'<5 Å) using the native sequence. **e,** Example structures predicted by AlphaFold3 using sequences generated by InvDNA, ViennaRNA and NUPACK for a G-quadruplex backbone (PDB: 8D79). All predicted structures are displayed in orientations aligned with experimental backbone. **f,** RMSD-C3' of structures predicted using InvDNA-generated sequences decreases as the number of generated sequence increase. Each sequence is generated using flexible backbone representation with either Gaussian noise (standard deviation=0.02 Å) or without noise (standard deviation=0.00 Å). **g,** As the number of sequences generated by InvDNA increases, the number of non-redundant sequences (pairwise sequence identity <100%) also increases.

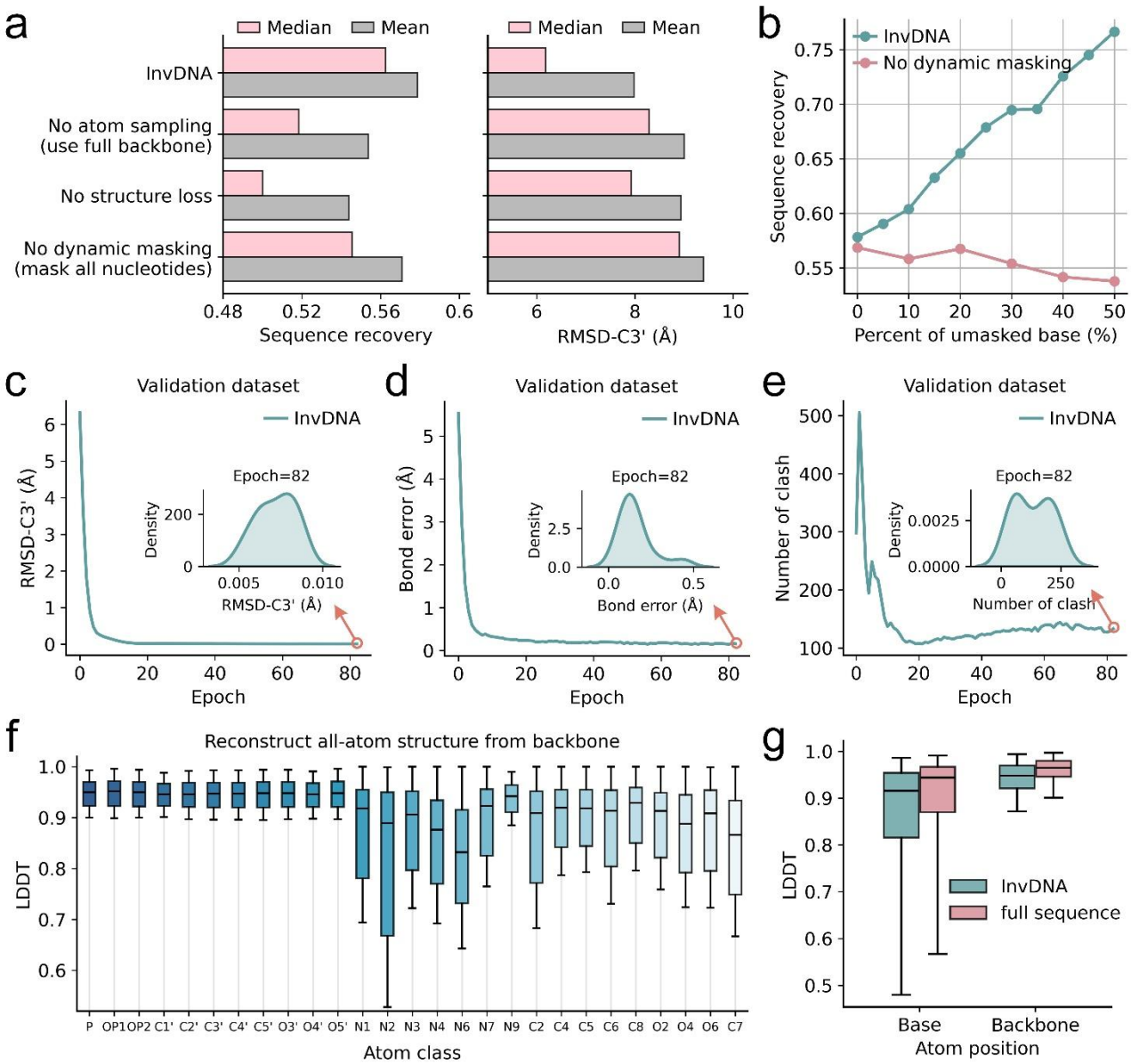

Figure 3: Ablation study of the training strategies in InvDNA. a, Mean and Median sequence recovery rates for sequences generated by InvDNA and by ablated model variants on the ssDNA dataset. Structures predicted from these generated sequences are also evaluated using RMSD-C3'. **b,** Sequence recovery rates for sequences generated by InvDNA and by the "No dynamic masking" model as the proportion of randomly retained nucleotides is gradually increased. **c-e,** Training curves for InvDNA on the validation dataset: **c:** RMSD-C3'; **d:** bond-length error; **e:** Number of atomic clashes. **f,** LDDT values for atoms in structures generated by InvDNA when full backbone atoms and the full sequence are provided as input. **g,** Distribution of LDDT for base atoms and backbone atoms in

structures generated by InvDNA and “full sequence” model when using full backbone atoms and full sequence as input.

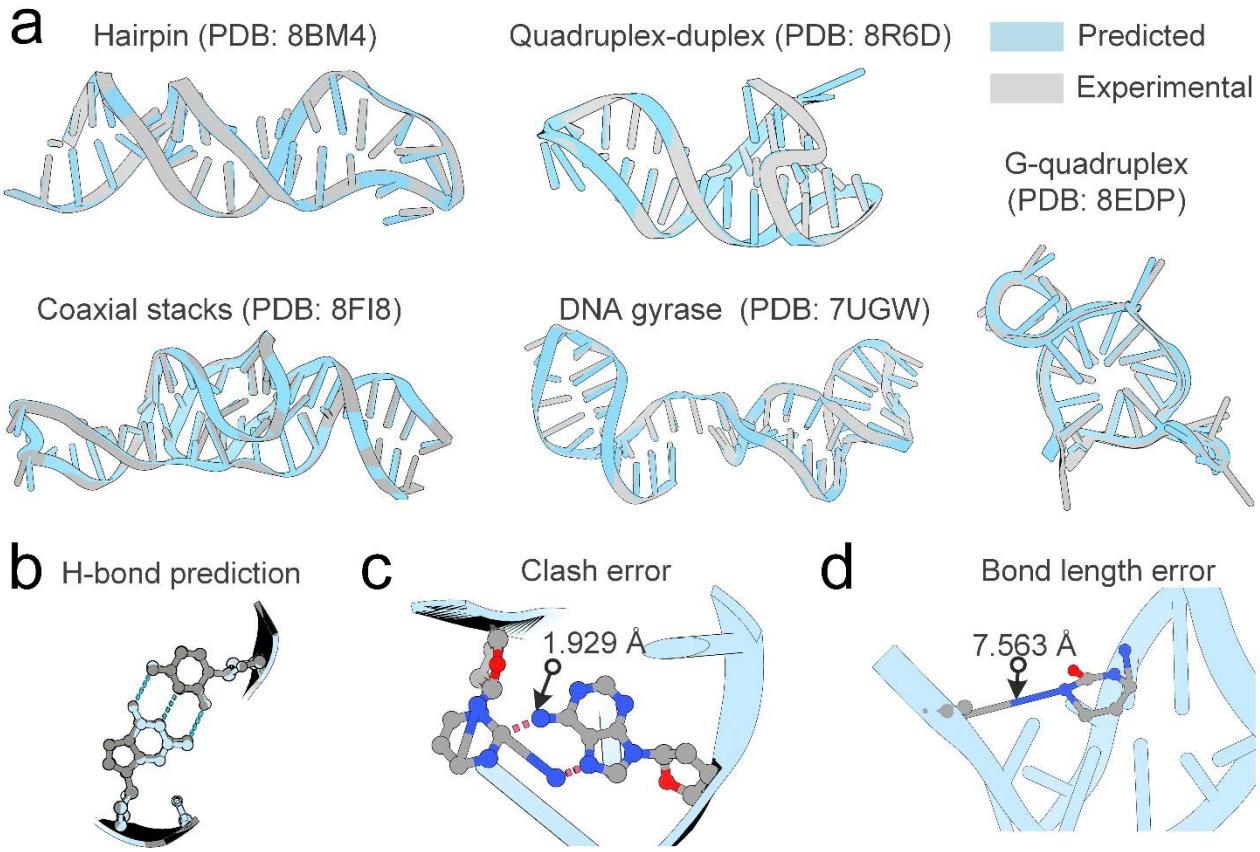

Figure 4: Examples of ssDNA structures predicted by InvDNA. a, The experimental structures are shown in grey, and the structures generated by InvDNA, using full backbone atoms and the full sequence as input, are shown in cyan. The nucleotides are depicted as stubs. **b-d,** Local details of the predicted structure: **b:** correctly predicted atomic positions and hydrogen bonds; **c:** atomic clash error; **d:** Bond-length error. All structures generated by InvDNA are available in the supplementary materials.

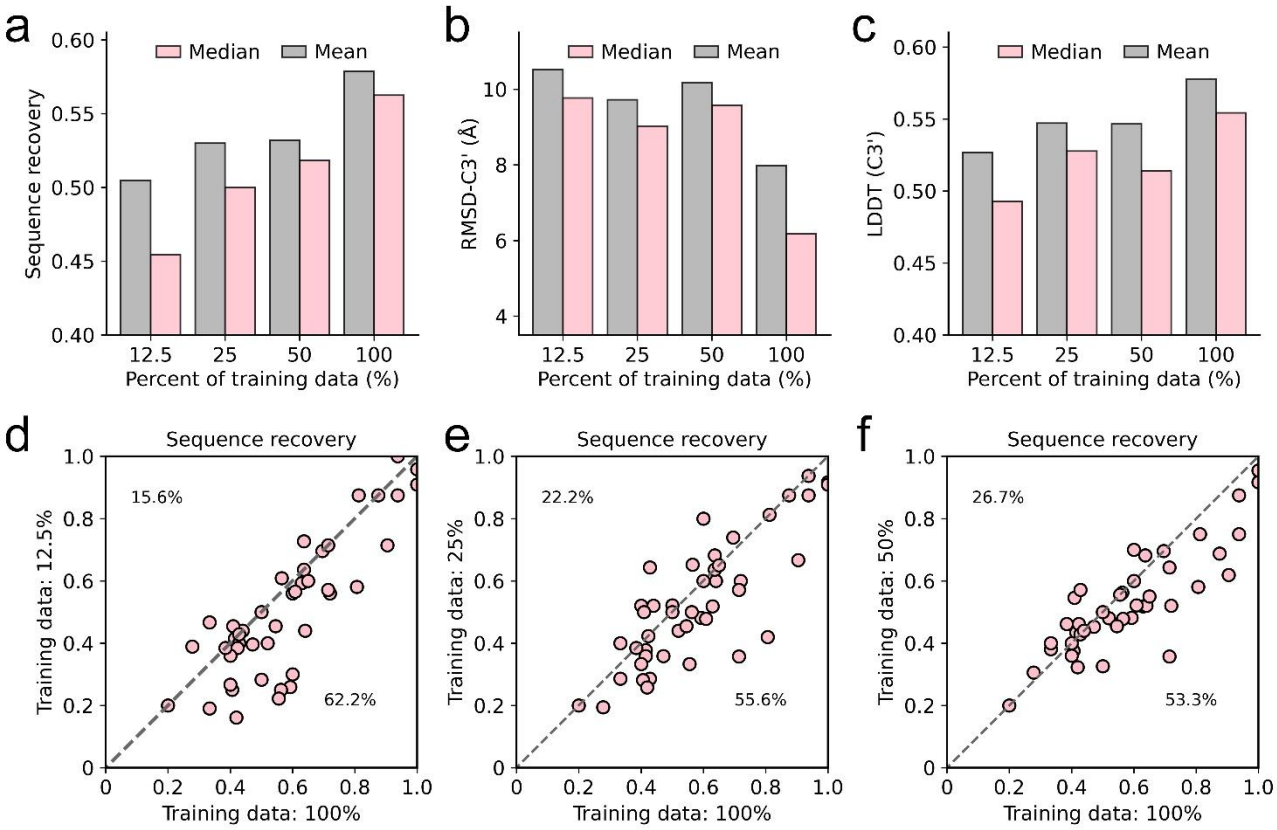

Figure 5: Training data analysis. a, Mean and Median sequence recovery rates for sequences generated by InvDNA and by ablated model variants on the ssDNA dataset. **b-c,** Structures predicted from these generated sequences are also evaluated using RMSD-C3' (**b**) and LDDT of C3' atoms (**c**). **d-f,** Pairwise comparisons between sequence generated by InvDNA (y-axis, Training data: 100%) and those generated by ablated model variants: **d,** Training data: 12.5%; **e,** Training data: 25%; **f,** Training data: 50%, on the ssDNA dataset.

Supplementary Information

Figures

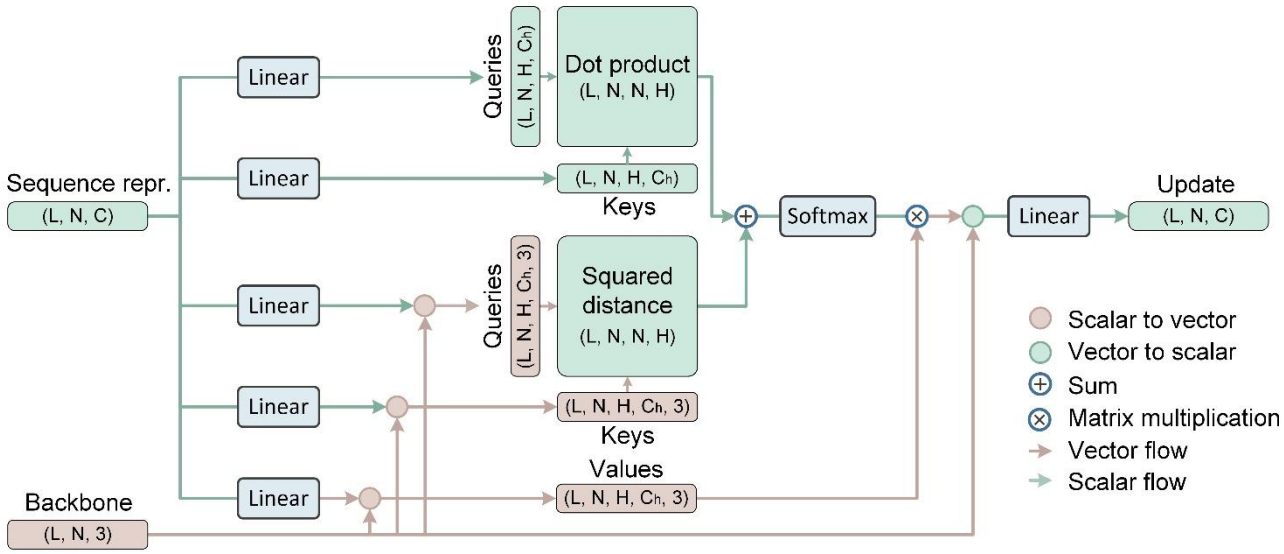

Figure S1: Architecture of the intra-nucleotide attention module. Linear refer to linear layer in neural network. The module takes sequence representation and backbone coordinates as input and outputs an updated sequence representation by learning intra-nucleotide interactions. Here, N denotes the number of atoms, L the sequence length, C the number of channels (default 384), and H the number of attention heads (default 12).

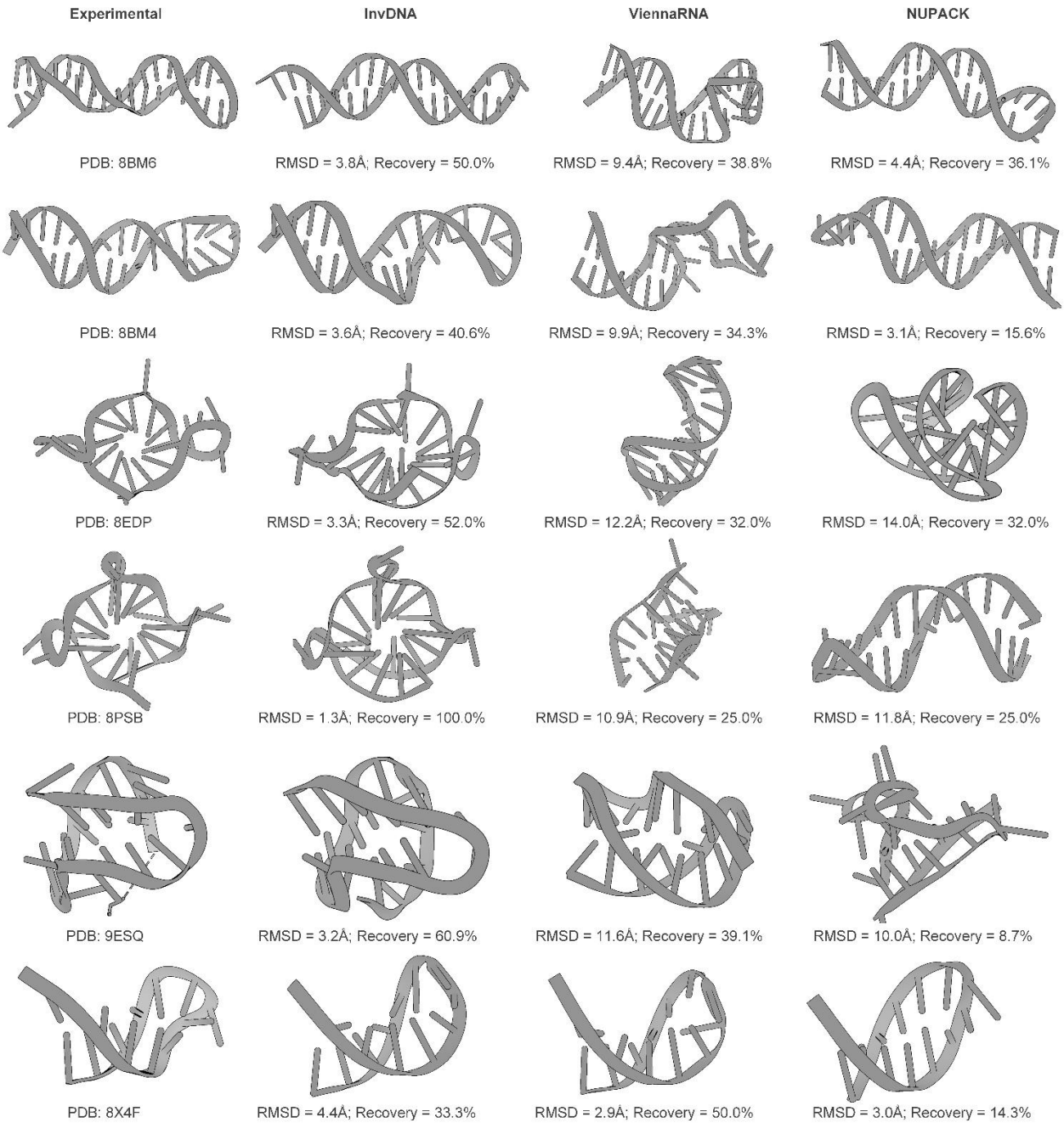

Figure S2: Examples of structures predicted by AlphaFold3 from sequences designed using InvDNA, ViennaRNA and NUPACK. In each row, the left panel shows the target backbone used as the template for sequence design, and all predicted structures are displayed in orientations aligned with their corresponding backbones. All sequences generated by three methods and all structures predicted by AlphaFold3 are available in the supplementary materials.

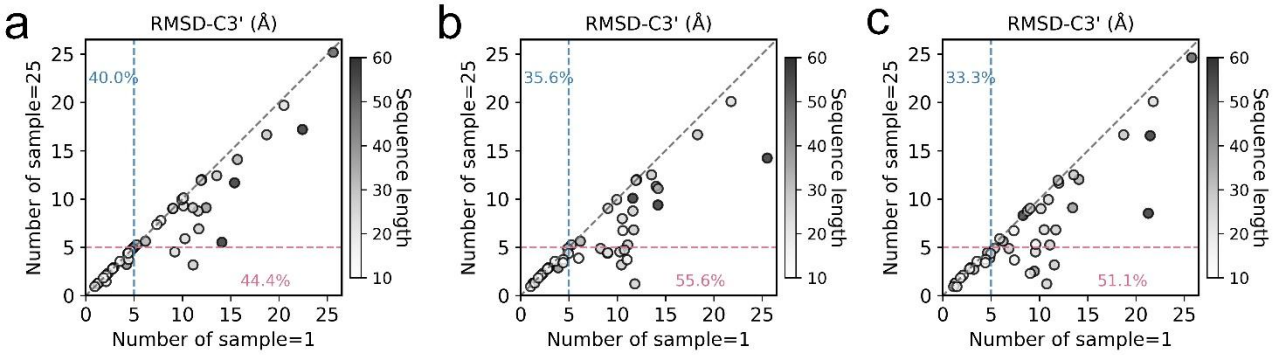

Figure S3: Pairwise comparisons between the best-performing predicted structures using InvDNA-generated sequences (x-axis: 1 generated sequence per backbone; y-axis: 25 generated sequences per backbone) on the ssDNA dataset. Each sequence is generated using a flexible backbone representation with either Gaussian noise (**b:** standard deviation=0.02 Å; **c:** standard deviation=0.04 Å) or without noise (**a:** standard deviation=0.00 Å).

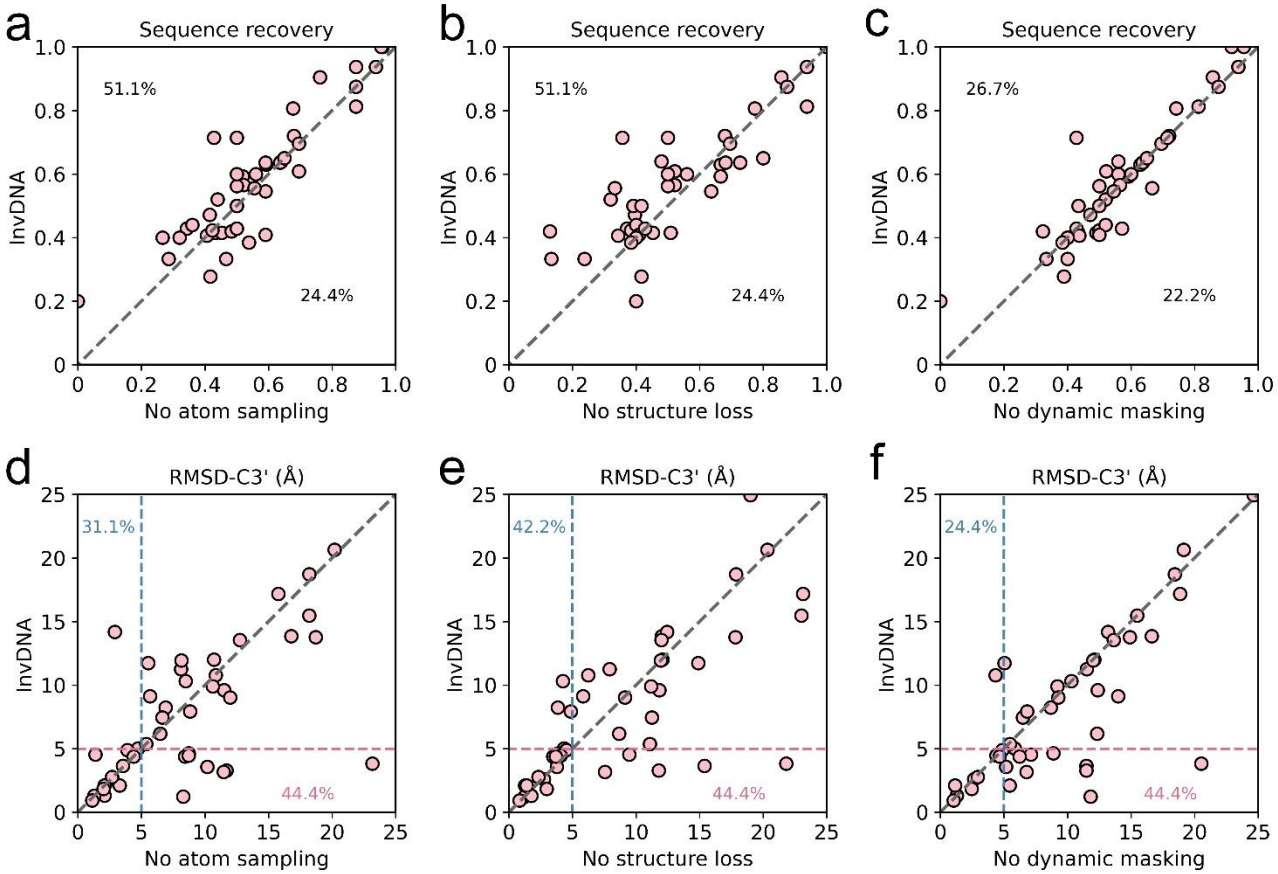

Figure S4: a-c, Pairwise comparisons between sequence generated by InvDNA (y-axis, Training data: 100%) and those generated by ablated model variants: **a,** "No atom sampling" model; **b,** "No structure loss" model; **c,** "No dynamic masking" model, on the ssDNA dataset. **d-f,** Pairwise comparisons between structures predicted using sequences generated by InvDNA (y-axis, Training data: 100%) and those generated by ablated model variants: **d,** "No atom sampling" model; **e,** "No structure loss" model; **f,** "No dynamic masking" model, on the ssDNA dataset. The structure is evaluated with RMSD-C3'.

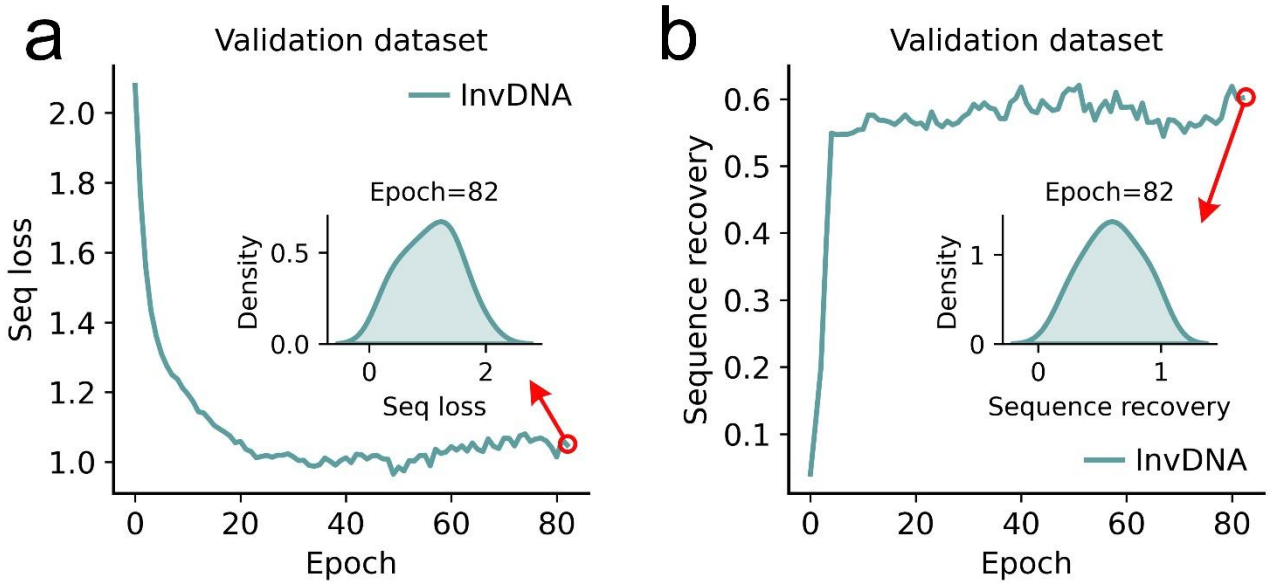

Figure S5: Training curves for InvDNA on the validation dataset: **a:** Mean sequence cross-entropy loss; **d:** Mean sequence recovery rate.

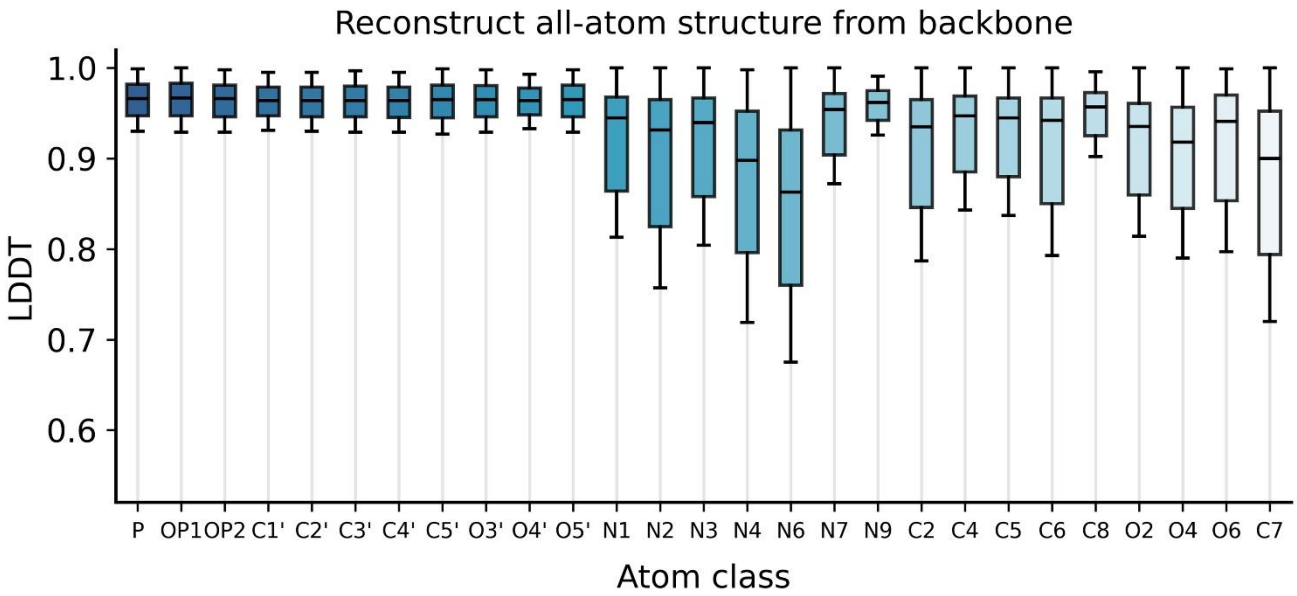

Figure S6: LDDT values for atoms in structures generated by "full sequence" model when full backbone atoms and the full sequence are provided as input.

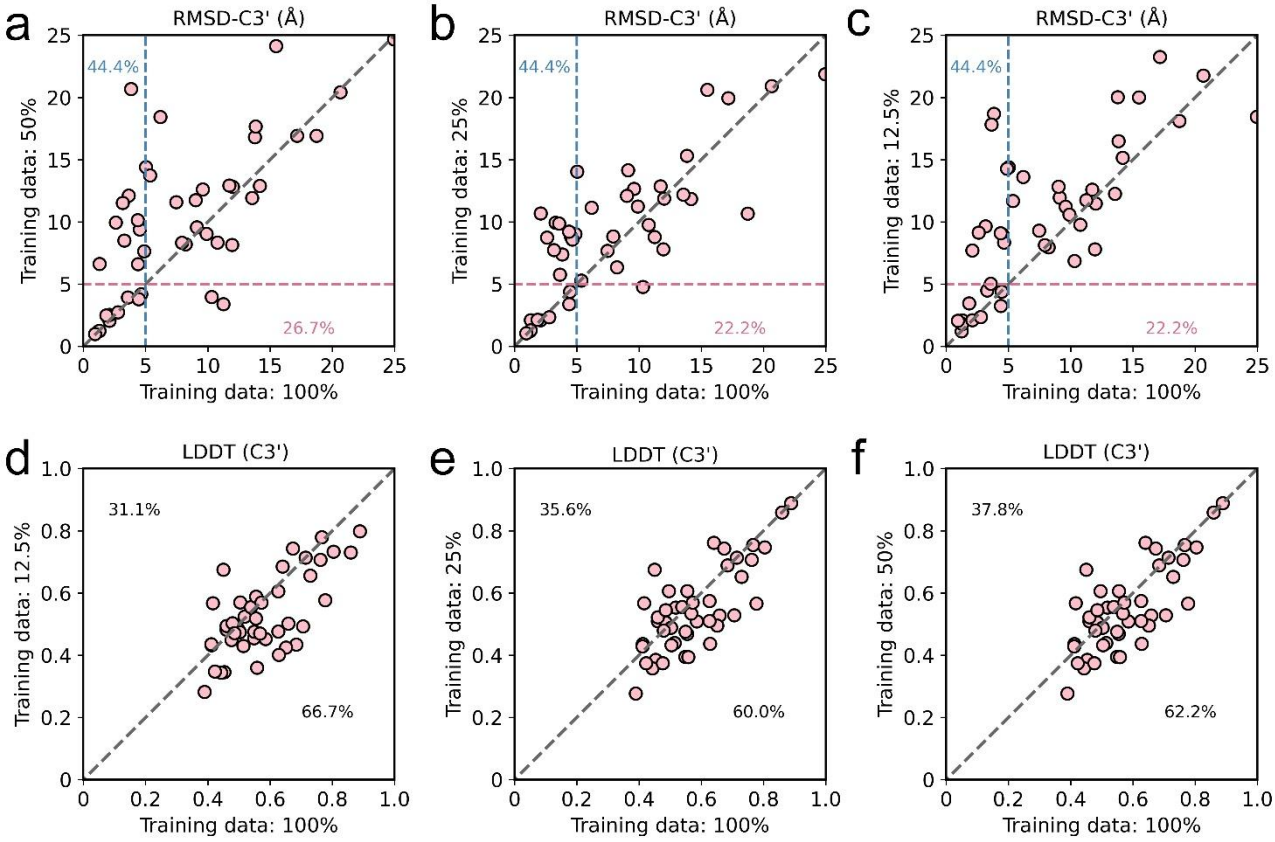

Figure S7: a-c, Pairwise comparisons between structures predicted using sequences generated by InvDNA (y-axis, Training data: 100%) and those generated by ablated model variants: **a,** Training data: 12.5%; **b,** Training data: 25%; **c,** Training data: 50%, on the ssDNA dataset. The structure is evaluated with RMSD-C3'. **d-f,** Pairwise comparisons between structures predicted using sequences generated by InvDNA (y-axis, Training data: 100%) and those generated by ablated model variants: **d,** training data: 12.5%; **e,** training data: 25%; **f,** training data: 50%, on the ssDNA dataset. The structure is evaluated with LDDT of C3' atoms.

Tables

Table S1: Performance of ablation models used in "Interpreting the InvDNA" on the ssDNA dataset

Model	Sequence recovery		RMSD (Å)	
	Median	Mean	Median	Mean
InvDNA	56.3%	57.9%	6.18	7.98
No atom sampling	51.9%	55.4%	8.29	9.01
No structure loss	50.0%	54.3%	7.92	8.94
No dynamic masking	54.5%	57.1%	8.90	9.39

Table S2: Performance of ablation models used in "Training data size analysis" on the ssDNA dataset

Training data	Sequence recovery		RMSD (Å)		LDDT	
	Median	Mean	Median	Mean	Median	Mean
12.5%	45.4%	50.5%	9.77	10.5	0.493	0.527
25%	50.0%	53.0%	9.03	9.72	0.528	0.547
50%	51.8%	53.2%	9.57	10.2	0.514	0.547
100%	56.3%	57.9%	6.18	7.98	0.554	0.578