Synthetic and Systems Biotechnology 10 (2025) 442–451

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: www.keaipublishing.com/en/journals/synthetic-and-systems-biotechnology

Original Research Article

Design and structure of overlapping regions in PCA via deep learning

Yan Zheng [a,b,1], Xi-Chen Cui [a,b,1], Fei Guo [a,c], Ming-Liang Dou [a], Ze-Xiong Xie [a,b,**], Ying-Jin Yuan [a,b,*]

[a] Frontiers Science Center for Synthetic Biology and Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin, 300072, PR China
[b] School of Chemical Engineering and Technology, Tianjin University, Tianjin, 300072, PR China
[c] School of Computer Science and Engineering, Central South University, Changsha, 410083, PR China

ARTICLE INFO

Keywords:
Synthetic biology
PCA
Deep learning
Molecular dynamics

ABSTRACT

Polymerase cycling assembly (PCA) stands out as the predominant method in the synthesis of kilobase-length DNA fragments. The design of overlapping regions is the core factor affecting the success rate of synthesis. However, there still exists DNA sequences that are challenging to design and construct in the genome synthesis. Here we proposed a deep learning model based on extensive synthesis data to discern latent sequence representations in overlapping regions with an AUPR of 0.805. Utilizing the model, we developed the SmartCut algorithm aimed at designing oligonucleotides and enhancing the success rate of PCA experiments. This algorithm was successfully applied to sequences with diverse synthesis constraints, 80.4 % of which were synthesized in a single round. We further discovered structure differences represented by major groove width, stagger, slide, and centroid distance between overlapping and non-overlapping regions, which elucidated the model's reasonableness through the lens of physical chemistry. This comprehensive approach facilitates streamlined and efficient investigations into the genome synthesis.

1. Introduction

The *de novo* synthesis of DNA sequences serves as a cornerstone in biology. Rapid DNA synthesis is a fundamental technique empowering scientists and engineers to discover and direct the basic activities of cells and organisms, thus driving advancements in biology across diverse fields [1].

Polymerase cycling assembly (PCA) stands out as the predominant method for the *de novo* synthesis of DNA fragments [2]. It was employed in the chemical synthesis from genes [2], gene clusters [3] to even designer chromosomes and genomes [3–5]. Over the past two decades, PCA has demonstrated its simplicity and efficacy in the construction of synthetic genomes spanning from small viral genomes in kilobases [6] to larger eukaryotic chromosomes in megabases [7]. It has been widely used in the first synthetic eukaryote genome, Sc2.0 (the Synthetic Yeast Genome Project) [8–10]. These achievements in synthetic biology have ushered in new opportunities for research in directed evolution [11], disease modeling [12], and the DNA storage of extensive data [13,14]. The international Genome Project-write (GP-write) consortium envisions the synthetic genomes of higher animals and plants, which will challenge longer and more intricate DNA sequences [15,16].

However, since the process of designing these oligos in PCA is tedious and confusing [17], and the detailed mechanism in this process remains unclear, there still exists genome sequences that are challenging to design and synthesize with PCA. In the synthesis of the *Saccharomyces cerevisiae* chromosome synV [18], certain building blocks necessitated multiple iterations involving trial-and-error adjustments of PCA experiments (Supplementary Information Table S1, Supplementary Information Section 1 and Supplementary Information Fig. S1). Similar difficulties arose during attempts to synthesize building blocks of *Caulobacter ethensis-1.0*. To overcome these challenges, large-scale sequence rewriting was implemented to facilitate the synthesis process, involving the substitutions of 10,172 bases and the removal of 5668 synthesis constraints [19]. While synonymous substitution could reduce synthesis

* Corresponding author. Frontiers Science Center for Synthetic Biology and Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin, 300072, PR China.
** Corresponding author. Frontiers Science Center for Synthetic Biology and Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin, 300072, PR China.
E-mail addresses: xzx@tju.edu.cn (Z.-X. Xie), yjyuan@tju.edu.cn (Y.-J. Yuan).
[1] Authors contributed equally to this work.

https://doi.org/10.1016/j.synbio.2024.12.007
Received 12 September 2024; Received in revised form 11 December 2024; Accepted 19 December 2024
Available online 27 December 2024

difficulty, its applicability is limited since the perfect matching of the assembled sequences to the design is crucial to verify design principles. As exploration extends into more intricate sequences, challenging sequences exerts a tangible impact on synthesis efficiency [20,21].

The precise pairing of overlapping regions assumes a pivotal role in the synthesis efficiency of PCA [22]. In the PCA process, the initial step involves the design and chemical synthesis of multiple oligonucleotides (oligos), typically ranging from 60 to 80 nucleotides in length. Overlapping regions, usually 15–25 nucleotides in length, are designed at the termini of adjacent oligos [22]. Thermodynamic collisions of overlapping regions lead to the pairing of complementary DNA molecules, a critical aspect for the uniqueness and correct hybridization of oligos. Following the formation of complementary base pairs in overlapping regions, DNA polymerase leverages the non-overlapping regions as a template to follow [23]. In this way, oligos are annealed and recursively elongated to generate the full-length DNA sequence (Supplementary Information Fig. S2). Despite the significance of overlapping regions, previous investigations have merely relied on empirical biological parameters [17,24] to design, avoiding extreme GC content and repetitive sequences. Thus, comprehensive research is expected to examine the sequence characteristics of overlapping regions through a data-driven approach instead of empirical parameters.

In this interdisciplinary study, we present a pioneering approach that integrates deep learning with the design of overlapping regions in PCA, aimed at optimizing DNA assembly (Fig. 1). We established a large synthesis dataset consisting of 32,714 PCA synthesized sequences and the corresponding design of overlapping regions. Based on the dataset, we then trained a deep learning model to discover the latent characteristics of overlapping regions. Utilizing this model, the SmartCut algorithm was developed to facilitate the rapid and accurate assembly of target constructs. The algorithm successfully designed and synthesized challenging sequences with diverse synthesis constraints in a single attempt. Furthermore, we extend our investigation into the structural aspects of overlapping regions through incorporating molecular dynamics (MD) simulations. By identifying subtle structure differences presented by parameters major groove width, stagger, slide, and centroid distance, our simulations contributed to a deeper understanding of the potential structural dynamics in the synthesis process, and examined the reasonableness of the model through a physical chemistry aspect.

2. Results

2.1. Modeling overlapping regions via deep learning

To discover the latent characteristics of overlapping regions, we collaborated with a gene synthesis company, Tsingke Biotech Co., Ltd. (www.tsingke.com) and collected a total of 32,714 DNA sequences. All of these sequences were successfully synthesized through PCA experiments, wherein designed overlapping regions were annotated. This strategy based on the fact that key design principles have been summarized in literature for overlapping regions, such as avoiding extreme GC content, secondary structures, repeats, etc. Thus, these characteristics will lead overlapping regions to be different from other sequences. This dataset encompassed sequences deriving from diverse species and featuring variations in both GC content and length (Methods). The diversity aimed to ensure its representativeness of sequences encountered during the synthesis process.

Systematic organization of oligo designs yielded 376,782 validated overlapping regions and 344,068 non-overlapping regions. Since

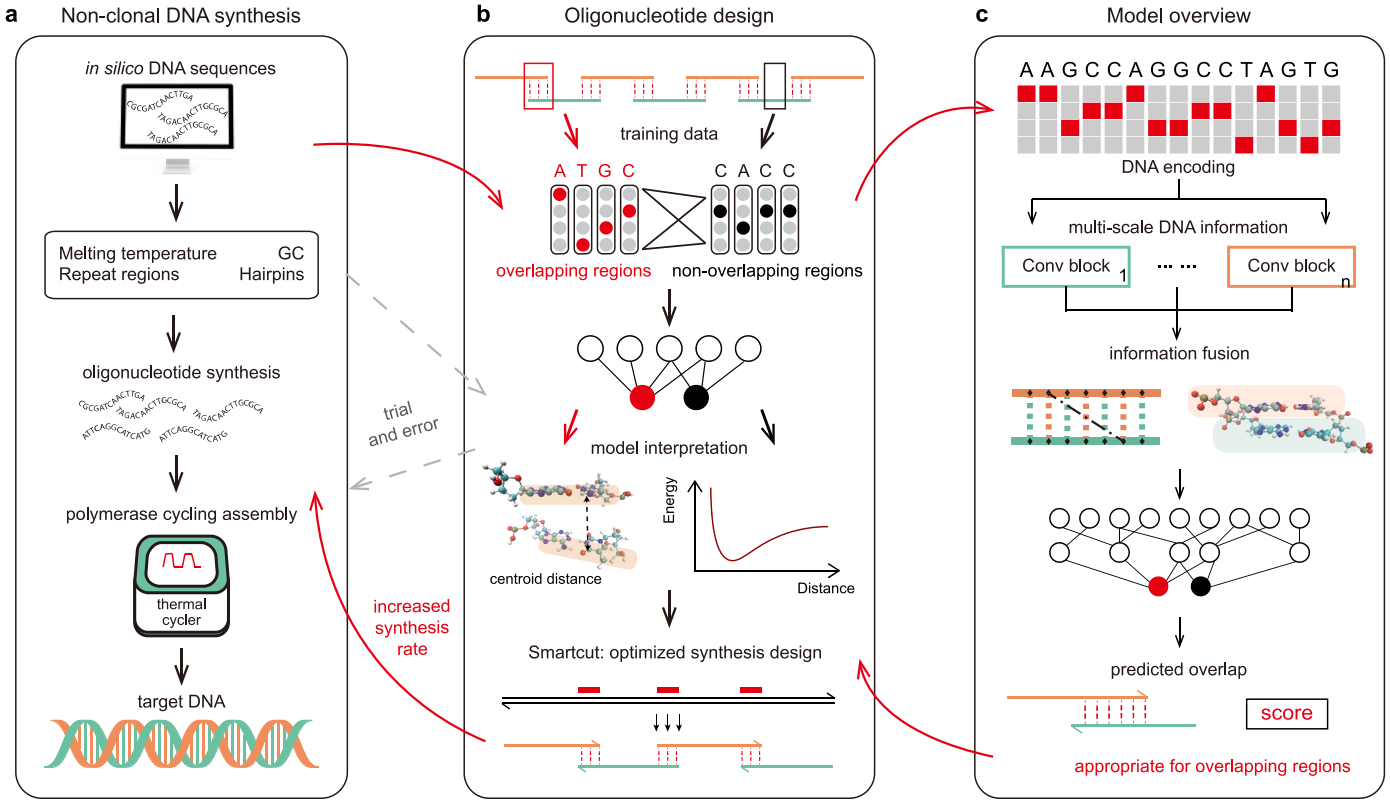


Fig. 1. Overview of design for DNA synthesis. (a) Non-clonal DNA synthesis. The synthesis initiates with the computational design of oligonucleotides, commonly taking into account various biological parameters. Subsequently, these oligonucleotides are assembled into the target DNA through PCA experiments. (b) Oligonucleotide design. Here we develop a new design algorithm, SmartCut. It differs from conventional approaches by utilizing a deep learning model to discern sequence representations of overlapping regions instead of relying on biological parameters. (c) Model overview. The deep learning model employed in SmartCut captures multi-scale DNA information, facilitating the evaluation of overlapping regions.

synthesis failures cannot always be attributed solely to specific overlapping regions, we used non-overlapping regions as decoy samples to help the model learn the distinguishing features of effective overlapping regions. This approach aligns with established practices in the literature, where negative decoy sequences are generated by random shuffling of subsequences from UniProt or positive sequences [25,26]. After exclusion criteria was used to eliminate sequences with high similarity, the final dataset comprised 41,334 overlapping and 40,207 non-overlapping sequences, randomly shuffled and partitioned into training, validation, and testing sets at a ratio of 3:1:1.

The proposed model architecture, the SmartCut model, integrates multiple convolutional blocks with distinct kernel sizes to capture information across diverse scales within overlapping regions. The model takes as input the One-hot encoding of DNA and it is trained to distinguish overlapping from non-overlapping regions. Grid search was performed to determine the optimal configuration of convolutional layers, activation functions, and other relevant layers. (Supplementary Information Section 2). Kernel sizes of convolutional layers, set as 3, 4, and 5, were selected to encompass both local base pair geometry parameters and a broader spectrum of structural information. Especially, the model aims at the distinctive energy signatures characterizing base pairs within the major groove, a region spanning 5 nucleobases. After convolutional blocks, the resulting latent space is subsequently concatenated for the classification of overlapping and non-overlapping regions using a prediction head, a linear layer used for classification.

The assessment of the efficacy of the SmartCut model involved a comparison with four architectures (Fig. 3a). A conventional two-layer convolutional neural network was first employed to align with the convolutional blocks in our model. Subsequently, we performed standard fine-tuning for DNABERT-2-117 M [27], a state-of-the-art large language model pre-trained on multi-species genomes. Further, we considered a two-layer Bidirectional Long Short-Term Memory

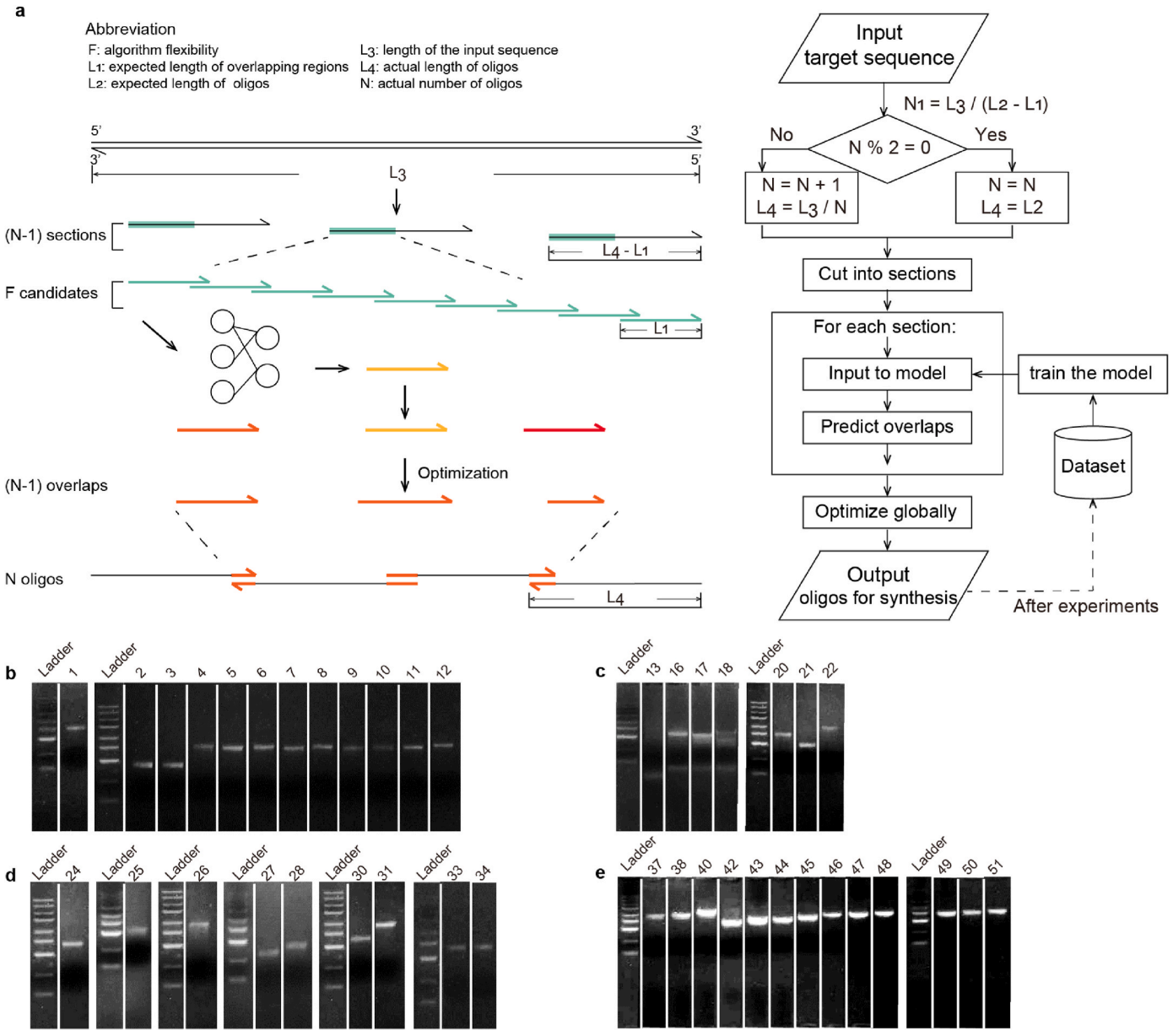


Fig. 2. Workflow and validation of the SmartCut algorithm. The SmartCut algorithm leverages a deep learning model to design oligos for synthesis (a). The input sequences undergo segmentation into sections and the model is employed to predict the optimal sequence for overlapping regions in each section. Following a global optimization process, the algorithm outputs the final set of oligos. The algorithm was successfully applied to challenging sequences with high GC (b, index 1–12), low GC (c, index 13–22), high S-index (d, index 23–36) and >1 kb length (e, index 37–51). See Github for detailed sequences. All the ladders were DL5,000 DNA Marker.

(BiLSTM) model and a self-attention-based model, both widely utilized in sequence data [28,29]. All comparison models underwent training (or finetuning for DNABERT-2-117 M) on the training dataset and subsequent hyperparameter-tuning on the validation dataset, alongside our SmartCut model. Ultimately, our SmartCut model demonstrated superior performance on the test dataset, achieving the highest Area Under the Curve (AUROC) in comparison to the other five models (Fig. 3a). The Area under the Precision-Recall Curve (AUPR), 0.849, suggests that the model could capture the latent characteristics of overlapping regions from the dataset (Supplementary Information Section 2).

The investigation of sequence motifs in overlapping regions was conducted utilizing the standard MEME protocol [30] (Fig. 3b and c). Although the model exhibits a propensity towards specific sequence configurations in the overlapping regions, discernible motifs were conspicuously absent in non-overlapping regions. It is challenging to comprehend the model's underlying rationale solely through motif analysis. A comprehensive analysis is required with the incorporation of structure and energy to delve into how the model assesses the stability and viability of diverse sequences in the PCA design.

2.2. Developing SmartCut algorithm for PCA design

Here we develop SmartCut, an algorithm tailored for the design of oligos in PCA experiments. The output of this algorithm comprises a set of oligos directly applicable in the synthesis of a specified target DNA sequence (Fig. 2a).

Primarily, the algorithm initiates by segmenting the input DNA sequence into an odd number of contiguous sections, where each section's length is determined by the expected length of the oligos minus the expected length of the overlapping regions. Subsequently, potential overlap candidates are excised from each section, and the algorithm's flexibility is utilized to define the model's input range, which is a user-customizable parameter that controls how many sequences need to be input into the model to determine an overlapping region. In contrast to previous PCA algorithms, SmartCut exhibits a simplified consideration of biological parameters, placing substantial emphasis on the capabilities of the deep learning model in the design. A higher algorithm flexibility affords the model greater latitude to function on larger regions. For each section, all candidates are inputted into the model and scored their probabilities of being suitable overlapping regions. The candidate with the highest score is retained as the preliminary overlapping region. Subsequently, an optimization process is applied to all selected preliminary regions, modifying these overlapping regions to achieve similar melting temperatures and uniform GC content. The sequence is then divided according to these overlapping regions to output the requisite oligos.

To evaluate the effectiveness of the algorithm, a collection of DNA sequences was reviewed for their challenging synthesis using PCA experiments. Sequences characterized by either excessively high or low GC content, as well as those exceeding 1 kb in length, have been deemed unsuitable for PCA synthesis [31]. The assessment of synthesis difficulties was further refined through the application of the S-index, a comprehensive metric that integrates both sequence and structural properties with an XGBoost model [21]. Sequences with an S-index surpassing 0.5 were considered as challenging. As delineated in Table 1, the collection comprised 12 sequences with GC content exceeding 0.65, 10 sequences with GC content below 0.30, 14 sequences with an S-index surpassing 0.5, and 15 sequences with lengths ranging from 1.0 kb to 1.9 kb (Supplementary Information Table S2).

These sequences were subjected to the SmartCut algorithm using a consistent configuration. The expected length of overlapping regions and oligos were set as 18 nt and 80 nt respectively. All the test sequences and corresponding design results of oligos were uploaded to the GitHub repository. Subsequently, all designed oligos underwent standardized synthesis procedures (Methods), enabling a rigorous assessment of the algorithm's performance across diverse synthesis constraints.

All the DNA sequences exhibiting high GC content were successfully assembled in a single round with distinct and clear bands in the agarose gel electrophoresis (Fig. 2b). For sequences with low GC content, merely three sequences failed with shallow target bands (Fig. 2c). And these problematic sequences were successfully synthesized after a second attempt with segmental amplification. 64.3 % sequences with an S-index greater than 0.5 were successfully synthesized (Fig. 2d), and 13 out of 15 longer sequences spanning 1.0 kb–1.9 kb were successfully assembled in the first attempt (Fig. 2e). All sequences that initially failed were also eventually synthesized by synthesizing partial fragments before final assembly or altering experimental conditions (Supplementary Information Table S3). Remarkably, these sequences included those that Tsingke Biotech Co., Ltd. failed to synthesize (13, 19, 21–23, 27–38) and 15 of them were successfully assembled in the first attempt after re-designed by SmartCut. Through Sanger sequencing, we confirmed that the 41 DNA sequences with bright bands were successfully synthesized. The sequencing data has been uploaded to GitHub. SmartCut algorithm adeptly designed oligos for sequences that were historically considered challenging to synthesize using PCA. The high success rate of 80.4 % underscored the superior performance of the SmartCut algorithm in addressing various synthesis challenges associated with PCA synthesis.

We have also run these experimental validation sequences with DNAWorks, a widely used and accessible oligo design algorithm [17]. DNAWorks is extremely time-consuming and it takes over 1000 min to

Table 1
Challenging DNA sequences for SmartCut validation.

Synthesis constraint	Number of Samples	Detail	Rate of Successful Assembly
High GC	12	GC content >0.65	12/12
Low GC	10	GC content <0.30	7/10
High S-index	14	S-index >0.5	9/14
Length	15	1.0 kb–1.9 kb	13/15
Total			41/51

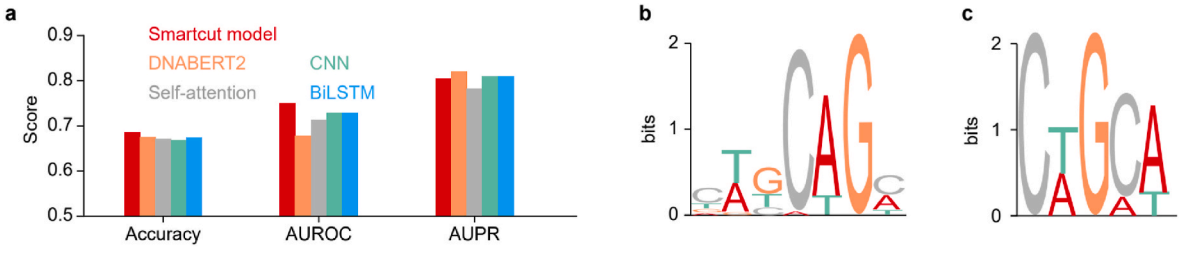


Fig. 3. Model comparison and motif discovery. (a) Model comparison with various widely-used models on the test dataset, as assessed through Area under the Accuracy, Area under Receiver Operator Curve (AUROC), and Area under the Precision-Recall Curve (AUPR). Motif of overlapped regions was calculated by MEME with the discriminative mode (b) and differential Enrichment mode (c), respectively.

handle 1700 bp target sequence, while SmartCut takes ~1 min for each sequence (Methods, Supplementary Information Fig. S4a). Furthermore, we performed a comparative analysis of Tm and GC content between design results of DNAWorks and SmartCut. We calculated the Tm of the deigned oligos using the primer3 package [32,33] and determined the difference between the highest Tm and lowest Tm for each target sequences (Supplementary Information Table S4). The average Tm difference of DNAWork is 15.34 °C, while it is only 9.30 °C for SmartCut, representing 0.61 times the variation observed with DNAWorks. The average GC content difference of DNAWork is 29 %, compared to 21 % for SmartCut, an 8 % reduction. These results demonstrate that SmartCut could design oligos with more uniform melting temperatures and GC content, thereby enhancing the success rate of PCA experiments.

To further demonstrate the practicality of SmartCut algorithm, we employed it to design building blocks for the synthetic yeast chromosome V and synthetic yeast chromosome X, the *Mycoplasma mycoides* JCVI-syn1.0 genome (GenBank: CP002027.1), and the *Escherichia coli* genome (NCBI Reference Sequence: NC_000913.3). Initially, we segmented the chromosomes and genomes into building blocks of 1 kb in length, appropriate for PCA experiments. Each building block was then processed with the SmartCut algorithm to design oligonucleotides and we recorded the time required to design all blocks for each chromosome or genome. The synthetic yeast chromosomes V and X, both under 1 Mb in length, were be completed within 3 min (Supplementary Information Fig. S4b). For the approximately 4 Mb *Escherichia coli* genome, the algorithm completed the design within 20 min. These results highlight the high computational efficiency of the SmartCut algorithm, confirming its suitability for genome synthesis projects.

2.3. Model interpretation based on dynamic simulation structures

To further explore the potential molecular mechanisms learned by the model, we subsequently investigated the differences in structure and energy between overlapping and non-overlapping regions utilizing model scores. We first designed a dataset consisting of representative sequences in the PCA experiments. Overlapping sequences were generated with a broad range of GC content percentages. Non-overlapping sequences, typically avoided in the overlapping region, were manually designed (Methods) and further refined through the previous experimental results and the expertise of frontline experimental personnel (Supplementary Information Table S5).

Next, we employed molecular dynamics simulations to investigate these DNA structures in solution environments under PCA experimental conditions (Supplementary Information Table S6) and examined the correlation between DNA structural characteristics and the model scores. We quantitatively characterized the structural parameters to conduct a comparative analysis between overlapping and non-overlapping molecules. Widely used base pair geometry parameters including shear, stretch, stagger, buckle, propeller, opening, shift, slide, rise, tilt, roll, and twist were examined, along with the widths of the minor and major grooves, and the centroid distance between bases (Supplementary Information Table S7 and Supplementary Information Fig. S6). To elucidate the factors influencing the success rates of PCA experiments, the Pearson correlation coefficient analysis was performed between the average values of parameters and SmartCut scores (Supplementary Information Table S8). Four parameters exhibiting relatively strong correlations ($|r| > 0.6$), major groove width, stagger (the translational parameter around the Z-axis of the dinucleotide intra-base pair), slide (the translational parameter around the Y-axis of the dinucleotide inter-base pair), and centroid distance (the Euclidean distance between the centroids of adjacent base pair planes), as depicted in Fig. 4a, were selected for further investigation.

The major groove width and centroid distance displayed negative correlations (Fig. 4b, major groove width $r = -0.650$, p-value <0.0001; Fig. 4e, centroid distance $r = -0.688$, p-value <0.0001), whereas the stagger and slide exhibited positive correlations (Fig. 4c, stagger $r = 0.694$, p-value <0.0001; Fig. 4d, slide $r = 0.700$, p-value <0.0001). As model scores increased, the averages of major groove width decreased to ~19.3 Å, stagger increased to ~0.12 Å, slide increased to ~ −0.7 Å, and centroid distance decreased to ~3.40 Å. Specifically, the centroid distance changed by 0.15 Å, the parameter slide changed by ~0.2 Å, and both major groove width and stagger changed by ~1 Å.

Kolmogorov-Smirnov tests revealed significant differences in all four parameter distributions between the two types of DNA molecules (Fig. 4f–i, p-value <0.0001). Furthermore, Kruskal-Wallis tests were performed to examine the differences in parameter distribution among pairwise molecules, and all four parameters exhibited statistically significant differences in pairs ($p < 0.0001$, Supplementary Information Table S9).

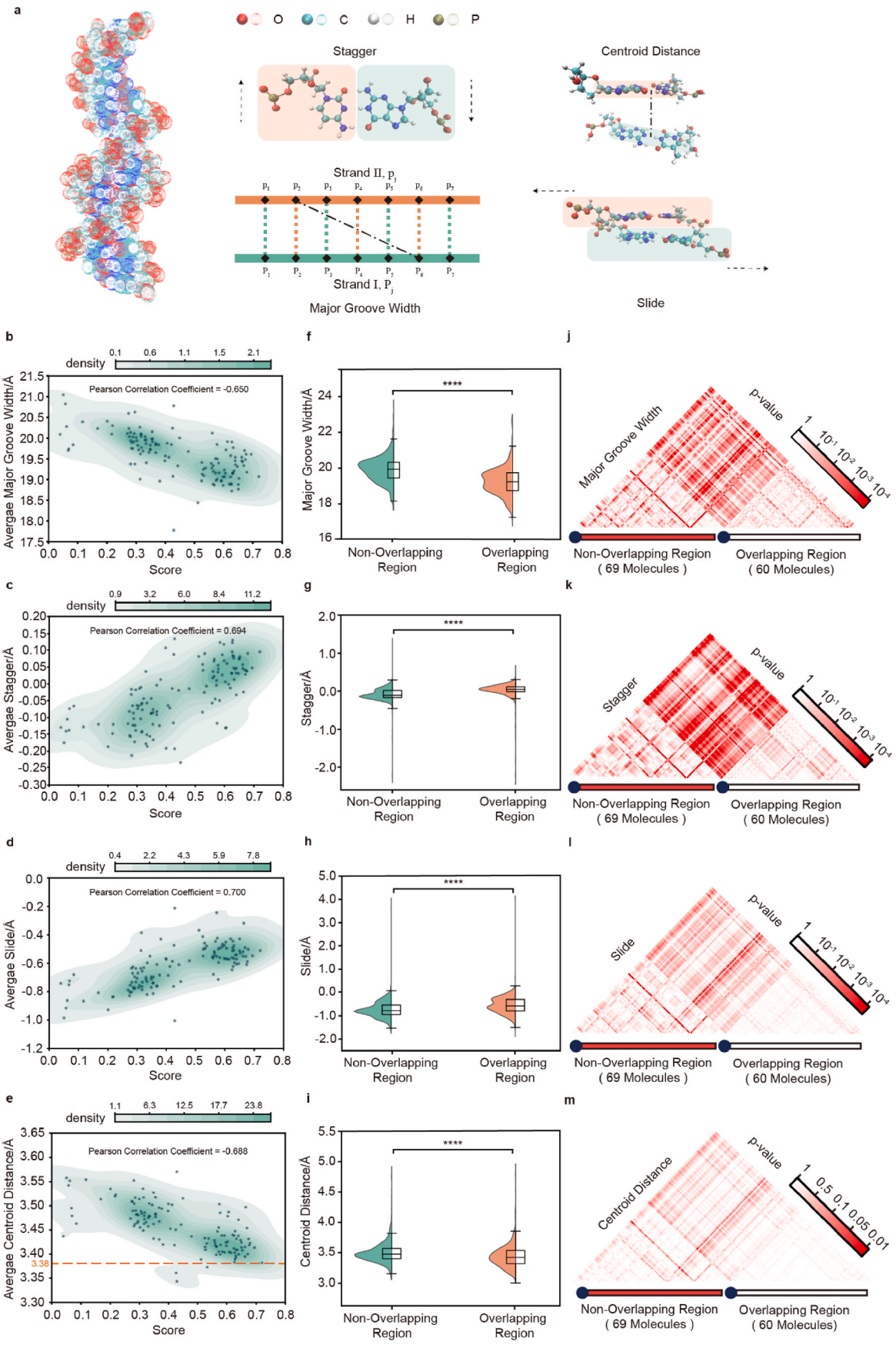
The post-hoc Dunn test with Benjamini/Hochberg correction generated p-value heat maps, which displayed the statistically significant differences in the distribution of these four parameters among all pairwise molecules (Fig. 4f–m). The intersection area (the rectangular area in the middle) exhibited significantly more pronounced differences than the individual areas of overlapping and non-overlapping molecules (the triangular area in the bottom right and bottom left), implying that major groove width, stagger, slide, and centroid distance are likely the discriminating factors between the two types of molecules. Notably, as the SmartCut scores increased, the average values of these four parameters of the molecules gradually approached the values of the parameters of standard B-form DNA molecules (The parameter slide is −0.33 Å, stagger is 0.01 Å, major groove width is 16.9 Å, and centroid distance is 3.38 Å in standard B-form DNA molecules, respectively).

To further investigate the biochemical nature behind the statistically differences and correlations, the centroid distance was selected as it is a key parameter that correlates with base stacking energy [34–36], which has been considered as one of the most important factors impacting the stability of DNA molecules [37–39]. We conducted rigid potential energy surface (PES) scans on the dinucleotides dimer by displacing a base along the stacking direction (Fig. 5a). Regardless of the scanning direction, the energy consistently increased compared to that of the standard B-form geometry. This indicates that standard B-form DNA molecule reached a minimum on the PES, which is usually the most stable structure in aqueous solutions. As mentioned earlier, the centroid distance decreased from ~3.55 Å to ~3.40 Å, gradually approaching the values of the parameters of the standard B-form DNA molecule. Correspondingly, the energy also dropped by ~14.12 kcal, suggesting that higher SmartCut scores lead to more stable DNA double strands, thereby enhancing the stability of the overlap binding section in the PCA process. To exclude the influence of other interactions, we conducted intermolecular interaction and energy decomposition analysis on the two bases extracted from the aforementioned structure, with the sugar backbone removed. The results also indicated an increase regardless of the scanning direction, with the intermolecular interaction decreasing by ~0.39 kcal. The decomposition analysis highlighted the contribution of pi-pi stacking interaction. The electrostatic interaction changed from stabilization to destabilization effect, while the sum of exchange, induction, and dispersion interaction changed from destabilization to stabilization effect. This led to the minimum of the intermolecular interaction at 3.38 Å, gradually diminishing as the centroid distance increased (Fig. 5b).

The analysis revealed that inherent physical chemistry features of functional overlapping sequences were extracted and underscored the reasonableness of the model from the perspective of structure and energy. Evaluated by actual PCA experiments, the model's proficiency in discerning stability characteristics serves as a reliable guide for the strategic design of oligos in the PCA process.

3. Discussion

In this study, we simplify the multi-variable problem of PCA design by focusing on the pivotal overlapping regions. A deep learning model



(caption on next page)

Fig. 4. Comparison and analysis of MD average structures of overlapping and non-overlapping regions. Comprehensive analysis of 129 molecules was conducted, including 69 non-overlapping regions and 60 overlapping regions. (a). Schematic plot of the four selected parameters with relatively strong correlations. The stagger parameter refers to the translational displacement along the Z-axis within the dinucleotide intra-base pair, while the slide parameter refers to the translational displacement along the Y-axis within the dinucleotide inter-base pair. The major groove width is depicted by a chain-dotted line in a flattened representation of eight base-pairs and seven phosphate groups on each DNA strand. Centroid distance refers to the Euclidean distance between the centroids of neighboring base pair planes. (b–e). Pearson correlation analysis between the average values of parameters and the SmartCut model scores. The dotted orange line refers to the 3.38 Å of parameter centroid distance. (f–i). The parameter distributions differences between two types of molecules performed by Kolmogorov-Smirnov test, the density scatter plot illustrates the kernel-density estimate with the colormap indicating the respective values. (j–m). The parameter distributions differences among pairwise molecules performed by Kruskal-Wallis test, the heat map plot displays the Benjamini/Hochberg corrected p-values of molecules in pairs.

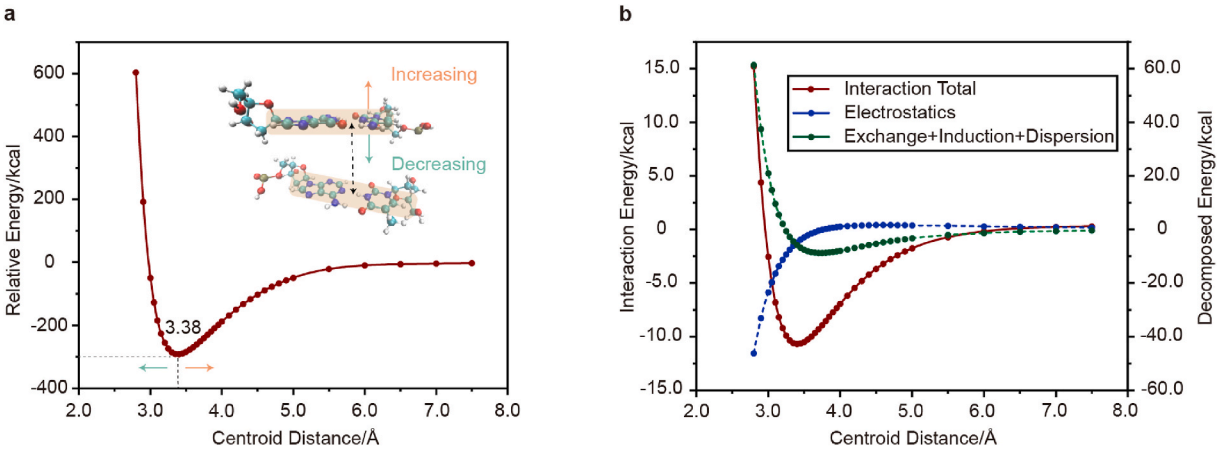


Fig. 5. Potential energy surface scan and interaction energy surface scan with energy decomposition analysis. Spacing scan between two dinucleotides at the ωB97X–V/def2-QZVPP level (a) and interaction energy scan and interaction components between base monomers in the dimer structure calculated at the scaled SAPT0/jun-cc-pVDZ level (b).

was first established to discern latent sequence characteristics distinguishing between overlapping and non-overlapping regions with multi-scale convolutional blocks. Subsequently, the predictive capability of this model was assimilated into the SmartCut algorithm, which designs oligos for the synthesis of target sequences. The algorithm's efficacy is substantiated through synthesizing sequences of heightened complexity, including extremely high and low GC content, high S-index and length exceeding 1 kb. To further explore the intrinsic mechanisms of the model, a series of MD experiments were devised to scrutinize the variations from both structural and energetic perspectives among diverse sequences.

Various oligo design methods emphasizing overlapping regions have been developed, including DNAWorks [17], GeneDesign [24,40], TmPrime [41], GeneGenie [42], and DropSynth [43,44]. These methods incorporate biological parameters of overlapping sequences such as GC content, repetitive regions, and melting temperature, which were derived from experiential summaries. While these design methods have contributed to improved efficiency in previous experiments, the intricate DNA interactions still introduce a propensity for errors in overlapping regions. It is challenging to rely on experiential parameters to elucidate complexities in structure and energy of overlapping regions. Mispairing, insertions, or deletions still lead to PCA failures [19,23]. These challenges underscore the requirement for a more comprehensive exploration of the intricate interplay within overlapping regions. While modifying experimental conditions can sometimes facilitate the synthesis of sequences involving difficult overlaps, the SmartCut algorithm offers a more systematic optimization approach. By addressing design issues from the outset, SmartCut minimizes the need for manual interventions and condition adjustments, which is particularly advantageous in high-throughput synthesis.

By deploying the data-driven deep learning model instead of biological parameters, our research not only augments the success rate of PCA experiments but also highlights the model's understanding of synthesis mechanics. The correlation between structural stability and the model score elucidates the rationale behind model predictions. Researches have previously explored the sequence-dependence of DNA structural stability [45–47]. However, the solution environments in these studies have not yet matched those in the actual PCA experiments. Our simulations set solution environments more closely resembling actual PCA experiments, thereby enhancing the credibility of the comparative analysis. Additionally, previous studies often explore combinations of relatively short DNA molecules with exhaustive computation, which is impractical since overlapping regions typically consist of at least 15 bp. The utilization of the SmartCut model allows for exploring representative sequences of overlapping regions in the PCA, facilitating a targeted investigation of sequence-dependent stability in longer DNA molecules. Notably, an increase in model scores correlates with a tendency for the DNA structure to approach the B-form. This finding aligns with previous studies that B-form DNA is more stable in low-salt solutions [48–50].

Despite these successes, certain limitations persist in our work. The algorithm overlooks the interactions among distinct overlapping regions in a sequence, and the influence of non-overlapping regions is not duly acknowledged on the assembly process. These deficiencies underscore opportunities for optimization in subsequent research endeavors. Due to the limitations on the length of synthesized polynucleotides, assembling oligos through overlapping regions is the main practical option for making large DNA sequences. Addressing these limitations will contribute to the continued advancement of the algorithm's efficacy and its applicability in the broader context of DNA sequence assembly methodologies.

4. Materials and methods

4.1. SmartCut dataset

The dataset encompassed both artificially designed and naturally derived sequences from diverse species, including *Escherichia coli*,

Saccharomyces cerevisiae, and *Homo sapiens*. The GC content of the sequences varied from 0.176 to 0.773, with a mean of 0.515, and the length ranged from 159 to 4600 base pairs (bp), with a mean of 865 bp. All of the synthesized sequences were designed using a modified version of DNAworks, with the melting temperature recommendations provided by the modified software. The synthesis was performed under identical experimental conditions at the company, following standard PCA procedures. From the oligo designs for all synthesized sequences, we collected 376,782 overlapping regions validated in the PCA experiments, which have contributed to the synthesis of sequences. Then we identified 344,068 non-overlapping regions which were positioned in the center of two overlapping regions, each sharing the average length of two adjacent overlapping regions. Consequently, the dataset consisted of 376,782 overlapping and 344,068 non-overlapping sequences. CD-HIT was used at the 80 % identity level to remove sequences bearing high similarity to others within the dataset.

4.2. Synthesis validation

For all the challenging sequences, we executed the SmartCut algorithm using the same configuration. The expected length of overlapping regions and oligos were set as 18 nt and 80 nt, respectively. The algorithm flexibility was 9. The oligos designed by the algorithm were sent to Synbio Technologies (https://synbio-tech.com/) for standard PCA synthesis. Experiments were carried out in 25 μL reaction mixtures containing 20 μL 1.25 × HiFi GS PCR Mix. Thermal cycling began with a 2-min denaturing step at 95 °C, and continued with 18 cycles at 95 °C for 15 s, 58 °C for 15 s, and 72 °C for 15 s, and finishing at 72 °C for 2 min.

4.3. Algorithm comparison

The performance of SmartCut was evaluated in comparison with popular oligo design programs. Due to the inaccessibility of the source codes for GeneDesign and TmPrime, this study focuses on comparing SmartCut with DNAWorks v3.2.4, a widely used and accessible oligo design algorithm for PCA experiments (https://github.com/davidhoover/DNAWorks). The DNAWorks program, which was last updated in 2017, was configured with specific parameters: a melting temperature lower bound of 62 °C using the "tolerance" argument, an oligonucleotide length lower bound of 75 nucleotides using the "random" argument, a frequency threshold of 10, sodium concentration set to 0.05 M, magnesium concentration set to 0.002 M, and the number of solutions set to 1. All other parameters were left at their default settings. The program was executed on a CentOS 7.5 system with 4 Intel Xeon 6348H processors (2.3 GHz, 24 cores).

4.4. MD simulations

In this study, we designed the non-overlapping sequences according to the following principles: (1) Extreme GC content (GC content <22 % or GC content >67 %); (2) High repeat density; (3) Formation of hairpins, loops, i-motifs, and G quadruplexes. A total of 129 DNA sequences were subjected to the SmartCut model, resulting in the categorization of 60 sequences as overlapping molecules and 69 as non-overlapping molecules.

The construction of initial B-form DNA duplexes with overlapping regions was carried out utilizing the University of California San Francisco Chimera program [51]. Subsequently, these initial structures were placed within rectangular boxes containing explicit water molecules and ions. To achieve system neutralization and alignment with practical PCA conditions, K+ counterions and KCl salt concentration of 0.264 mol/L were introduced. Following energy minimization and pre-equilibration of the systems, MD simulations lasting 20 ns were performed at a constant temperature of 313.15 K and pressure of 1 atm. These simulations employed periodic boundary conditions and the particle mesh Ewald method to manage long-range interactions, using a time step of 1 fs and the leap-frog algorithm. The DNA duplexes were allowed complete freedom of movement within the solution during these simulations. All Molecular Dynamics simulations, trajectory analyses, and time-average structure calculations were conducted using the GROMACS 2022.1 software package [52], employing the parmbsc1 force fields [53] and TIP3P water model [54] to account for molecular interactions. Parameters such as base-pair and base-pair step geometries, as well as measurements of minor and major groove widths, were determined using the x3dna package [55].

We extracted the time-averaged structure of the overlapping DNA duplex under simulated PCA experiment conditions by computing the root-mean-square deviations (RMSD) of trajectories. MD simulations of the initial B-form overlapping DNA duplex quickly reached equilibrium (Supplementary Information Fig. S5), indicating that 20 ns simulations were adequate due to the inherent stability of B-form DNA duplexes in aqueous environments (RMSD stabilized at ~3 Å).

4.5. QM calculations

The generation of initial structure of dinucleotides was conducted using the University of California San Francisco Chimera program, while the OpenBabel program [56] was utilized to neutralize the negative charges on the phosphate group through the addition of protons. Subsequently, the positions of the hydrogen atoms were optimized using the ORCA 5.0.3 program [57] employing the ωB97XD exchange-correlation functional [58] in conjunction with the def2-TZVP basis set [59], which has demonstrated satisfactory capabilities in describing various types of intermolecular interactions [60,61]. To obtain accurate single-point energy calculations for the molecular dimer, the ωB97X–V functional [62] in combination with a very large def2-QZVPP basis set [59] was employed. This combination has been demonstrated to provide excellent energy estimations for weakly interacting systems [63]. The RIJCOSX technique [64] was employed to expedite the computational calculations. Furthermore, Symmetry-Adapted Perturbation Theory (SAPT) analysis at the SAPT0 level, utilizing the jun-cc-pVDZ basis set, was conducted using the PSI4 1.9 code [65] to gain deeper insights into the binding energy. The SAPT input files for PSI4 and ORCA input files were generated with the assistance of the Multiwfn program [61]. The visualization of the molecular structures was achieved using the Visual Molecular Dynamics (VMD) software [66].

4.6. Statistics indicator

To examine potential differences between overlap and non-overlap structural parameters, and to evaluate correlations between these parameters and model scores, a thorough statistical analysis was performed. This analysis included the utilization of the Kolmogorov-Smirnov test, Kruskal-Wallis test, post hoc pairwise comparison test for multiple mean rank sum comparisons (Dunn's test) with Benjamini/Hochberg correction, and Pearson correlation coefficient calculations. These statistical procedures were executed using the scipy.stats package [67], following established analytical protocols. The selection of these statistical tests was based on well-established analytical guidelines. The Kolmogorov-Smirnov test statistic is expressed as follows:

$$D_n = \sup_x |F_{1n}(x) - F_{2n}(x)|$$

Where F_{n1} and F_{n2} are the empirical CDFs of the two samples and x is a point in the support of the data. The test statistic D_n is a measure of the maximum difference between the two CDFs. The empirical distribution function F_n for n independent and identically distributed ordered observations X_i is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i)$$

Where $1_{(-\infty,x]}(X_i)$ is the indicator function, equal to 1 if $X_i \leq x$ and equal

to 0 otherwise.

The Kruskal-Wallis test statistic is given by

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{r}_i - \bar{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(r_{ij} - \bar{r})^2}$$

Where N is the total number of observations across all groups, g is the number of groups, n_i is the number of observations in group i, r_{ij} is the rank (among all observations) of observation j from group i, $\bar{r}_i$ is the average rank of all observations in group i, $\bar{r}$ is the average of all the r_{ij}.

The Pearson correlation coefficient is given by

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where n is sample size, x_i, y_i are the individual sample points indexed with i, $\bar{x}$ and $\bar{y}$ are the sample means. The t-score of a correlation coefficient is given by

$$t = r_{ij}\frac{n-2}{1-r^2}$$

CRediT authorship contribution statement

Yan Zheng: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Xi-Chen Cui:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Fei Guo:** Writing – review & editing, Supervision, Conceptualization. **Ming-Liang Dou:** Formal analysis, Data curation, Conceptualization. **Ze-Xiong Xie:** Writing – review & editing, Supervision, Resources, Investigation, Funding acquisition, Data curation, Conceptualization. **Ying-Jin Yuan:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Formal analysis, Conceptualization.

Availability of data and materials

This work was developed in Python 3.8 using the Anaconda scientific computing stack. The data and code that support this work is available at https://github.com/zyan-y/SmartCut.

The DNA synthesis data used for model construction originates from Tsingke Biotech Co., Ltd. The data includes sequences intended for research or commercial purposes. Therefore, researchers wishing to obtain the data need contact Tsingke and sign the material transfer agreement to confirm that the data will be used solely for training the assembly model before access can be granted.

Funding

This work was funded by the National Key Research and Development Program of China (2022YFC2106300) and the National Natural Science Foundation of China (22378307).

Declaration of competing interest

A patent (202310487247.X) has been filed for the SmartCut algorithm presented in this study. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.synbio.2024.12.007.

References

[1] Aurand E, Keasling J, Friedman D, Salis H. Engineering Biology: a research roadmap for the next-generation bioeconomy. Engineering Biology Research Corporation, Emeryville, CA 2019.
[2] Hoose A, Vellacott R, Storch M, Freemont PS, Ryadnov MG. DNA synthesis technologies to close the gene writing gap. Nat Rev Chem 2023;7(3):144–61.
[3] Stemmer WP, Crameri A, Ha KD, Brennan TM, Heyneker HL. Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. Gene 1995;164(1):49–53.
[4] Annaluru N, Muller H, Mitchell LA, Ramalingam S, Stracquadanio G, Richardson SM, Dymond JS, Kuang Z, Scheifele LZ, Cooper EM, Cai Y, Zeller K, Agmon N, Han JS, Hadjithomas M, Tullman J, Caravelli K, Cirelli K, Guo Z, London V, Yeluru A, Murugan S, Kandavelou K, Agier N, Fischer G, Yang K, Martin JA, Bilgel M, Bohutski P, Boulier KM, Capaldo BJ, Chang J, Charoen K, Choi WJ, Deng P, DiCarlo JE, Doong J, Dunn J, Feinberg JI, Fernandez C, Floria CE, Gladowski D, Hadidi P, Ishizuka I, Jabbari J, Lau CY, Lee PA, Li S, Lin D, Linder ME, Ling J, Liu J, Liu J, London M, Ma H, Mao J, McDade JE, McMillan A, Moore AM, Oh WC, Ouyang Y, Patel R, Paul M, Paulsen LC, Qiu J, Rhee A, Rubashkin MG, Soh IY, Sotuyo NE, Srinivas V, Suarez A, Wong A, Wong R, Xie WR, Xu Y, Yu AT, Koszul R, Bader JS, Boeke JD, Chandrasegaran S. Total synthesis of a functional designer eukaryotic chromosome. Science 2014;344(6179):55–8.
[5] Kodumal SJ, Patel KG, Reid R, Menzella HG, Welch M, Santi DV. Total synthesis of long DNA sequences: synthesis of a contiguous 32-kb polyketide synthase gene cluster. In: Proceedings of the national academy of sciences of the United States of America, vol. 101; 2004. p. 15573–155738. 44.
[6] Smith HO, Hutchison CA, Pfannkoch C, Venter JC. Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. In: Proceedings of the national academy of sciences of the United States of America, vol. 100; 2003. p. 15440–5. 26.
[7] Wu Y, Li BZ, Zhao M, Mitchell LA, Xie ZX, Lin QH, Wang X, Xiao WH, Wang Y, Zhou X, Liu H, Li X, Ding MZ, Liu D, Zhang L, Liu BL, Wu XL, Li FF, Dong XT, Jia B, Zhang WZ, Jiang GZ, Liu Y, Bai X, Song TQ, Chen Y, Zhou SJ, Zhu RY, Gao F, Kuang Z, Wang X, Shen M, Yang K, Stracquadanio G, Richardson SM, Lin Y, Wang L, Walker R, Luo Y, Ma PS, Yang H, Cai Y, Dai J, Bader JS, Boeke JD, Yuan YJ. Bug mapping and fitness testing of chemically synthesized chromosome X. Science 2017;355(6329):eaaf4706.
[8] Cooper EM, Müller H, Chandrasegaran S, Bader JS, Boeke JD. The build-a-genome course. In: Peccoud J, editor. Gene synthesis: methods and protocols. Totowa, NJ: Humana Press; 2012. p. 273–83.
[9] Lauer S, Luo J, Lazar-Stefanita L, Zhang W, McCulloch LH, Fanfani V, Lobzaev E, Haase MAB, Easo N, Zhao Y, Yu F, Cai J, Build AGC, Bader JS, Stracquadanio G, Boeke JD. Context-dependent neocentromere activity in synthetic yeast chromosome VIII. Cell Genomics 2023;3(11):100437.
[10] McCulloch LH, Sambasivam V, Hughes AL, Annaluru N, Ramalingam S, Fanfani V, Lobzaev E, Mitchell LA, Cai J, Build AGC, Jiang H, LaCava J, Taylor MS, Bishai WR, Stracquadanio G, Steinmetz LM, Bader JS, Zhang W, Boeke JD, Chandrasegaran S. Consequences of a telomerase-related fitness defect and chromosome substitution technology in yeast synIX strains. Cell Genomics 2023;3(11):100419.
[11] Zhou S, Wu Y, Zhao Y, Zhang Z, Jiang L, Liu L, Zhang Y, Tang J, Yuan Y-J. Dynamics of synthetic yeast chromosome evolution shaped by hierarchical chromatin organization. Natl Sci Rev 2023;10(5):nwad073.
[12] Mitchell LA, McCulloch LH, Pinglay S, Berger H, Bosco N, Brosh R, Bulajic M, Huang E, Hogan MS, Martin JA, Mazzoni EO, Davoli T, Maurano MT, Boeke JD. De novo assembly and delivery to mouse cells of a 101 kb functional human gene. Genetics 2021;218(1):iyab038.
[13] Chen W, Han M, Zhou J, Ge Q, Wang P, Zhang X, Zhu S, Song L, Yuan Y. An artificial chromosome for data storage. Natl Sci Rev 2021;8(5):nwab028.
[14] Zhou J, Zhang C, Wei R, Han M, Wang S, Yang K, Zhang L, Chen W, Wen M, Li C, Tao W, Yuan YJ. Exogenous artificial DNA forms chromatin structure with active transcription in yeast. Sci China Life Sci 2022;65(5):851–60.
[15] Boeke JD, Church G, Hessel A, Kelley NJ, Arkin A, Cai Y, Carlson R, Chakravarti A, Cornish VW, Holt L, Isaacs FJ, Kuiken T, Lajoie M, Lessor T, Lunshof J, Maurano MT, Mitchell LA, Rine J, Rosser S, Sanjana NE, Silver PA, Valle D, Wang H, Way JC, Yang L. GENOME ENGINEERING. The genome project-write. Science 2016;353(6295):126–7.
[16] Ostrov N, Beal J, Ellis T, Gordon DB, Karas BJ, Lee HH, Lenaghan SC, Schloss JA, Stracquadanio G, Trefzer A, Bader JS, Church GM, Coelho CM, Efcavitch JW, Guell M, Mitchell LA, Nielsen AAK, Peck B, Smith AC, Stewart Jr CN, Tekotte H. Technological challenges and milestones for writing genomes. Science 2019;366(6463):310–2.
[17] Hoover DM, Lubkowski J. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. Nucleic Acids Res 2002;30(10):e43.
[18] Xie ZX, Li BZ, Mitchell LA, Wu Y, Qi X, Jin Z, Jia B, Wang X, Zeng BX, Liu HM, Wu XL, Feng Q, Zhang WZ, Liu W, Ding MZ, Li X, Zhao GR, Qiao JJ, Cheng JS, Zhao M, Kuang Z, Wang X, Martin JA, Stracquadanio G, Yang K, Bai X, Zhao J, Hu ML, Lin QH, Zhang WQ, Shen MH, Chen S, Su W, Wang EX, Guo R, Zhai F, Guo XJ, Du HX, Zhu JQ, Song TQ, Dai JJ, Li FF, Jiang GZ, Han SL, Liu SY, Yu ZC, Yang XN, Chen K, Hu C, Li DS, Jia N, Liu Y, Wang LT, Wang S, Wei XT, Fu MQ, Qu LM, Xin SY, Liu T, Tian KR, Li XN, Zhang JH, Song LX, Liu JG, Lv JF, Xu H, Tao R, Wang Y, Zhang TT, Deng YX, Wang YR, Li T, Ye GX, Xu XR, Xia ZB, Zhang W, Yang SL, Liu YL, Ding WQ, Liu ZN, Zhu JQ, Liu NZ, Walker R, Luo Y, Wang Y, Shen Y, Yang H, Cai Y, Ma PS, Zhang CT, Bader JS, Boeke JD, Yuan YJ. "Perfect" designer chromosome V and behavior of a ring derivative. Science 2017;355(6329):eaaf4704.

[19] Venetz JE, Del Medico L, Wolfle A, Schachle P, Bucher Y, Appert D, Tschan F, Flores-Tinoco CE, van Kooten M, Guennoun R, Deutsch S, Christen M, Christen B. Chemical synthesis rewriting of a bacterial genome to achieve design flexibility and biological functionality. In: Proceedings of the national academy of sciences of the United States of America, vol. 116; 2019. p. 8070–9. 16.
[20] Christen M, Deutsch S, Christen B. Genome Calligrapher: A web tool for refactoring bacterial genome sequences for de Novo DNA synthesis. ACS Synth Biol 2015;4(8):927–34.
[21] Zheng Y, Song K, Xie ZX, Han MZ, Guo F, Yuan YJ. Machine learning-aided scoring of synthesis difficulties for designer chromosomes. Sci China Life Sci 2023;66(7):1615–25.
[22] Hughes RA, Ellington AD. Synthetic DNA synthesis and assembly: putting the synthetic in synthetic biology. Cold Spring Harbor Perspect Biol 2017;9(1):a023812.
[23] Xiong AS, Peng RH, Zhuang J, Gao F, Li Y, Cheng ZM, Yao QH. Chemical gene synthesis: strategies, softwares, error corrections, and applications. FEMS (Fed Eur Microbiol Soc) Microbiol Rev 2008;32(3):522–40.
[24] Richardson SM, Nunley PW, Yarrington RM, Boeke JD, Bader JS. GeneDesign 3.0 is an updated synthetic biology toolkit. Nucleic Acids Res 2010;38(8):2603–6.
[25] Samantha P, Sean RE. Constructing benchmark test sets for biological sequence analysis using independent set algorithms. PLoS Comput Biol 2022;18(3):e1009492.
[26] Babak A, Andrew D, Matthew TW, Brendan JF. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol 2015;33(2):831–8.
[27] Zhou Z, Ji Y, Li W, Dutta P, Davuluri R, Liu H. Dnabert-2: efficient foundation model and benchmark for multi-species genome. arXiv preprint arXiv:2306.15006 2023.
[28] Guan Y, Li H, Yi D, Zhang D, Yin C, Li K, Zhang P. A survival model generalized to regression learning algorithms. Nature Computational Science 2021;1(6):433–40.
[29] Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF, Wicky BIM, Hanikel N, Pellock SJ, Courbet A, Sheffler W, Wang J, Venkatesh P, Sappington I, Torres SV, Lauko A, De Bortoli V, Mathieu E, Ovchinnikov S, Barzilay R, Jaakkola TS, DiMaio F, Baek M, Baker D. De novo design of protein structure and function with RFdiffusion. Nature 2023;620(7976):1089–100.
[30] Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res 2015;43(W1):39–49.
[31] Czar MJ, Anderson JC, Bader JS, Peccoud J. Gene synthesis demystified. Trends Biotechnol 2009;27(2):63–72.
[32] Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics 2007;23(10):1289–91.
[33] Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3-new capabilities and interfaces. Nucleic Acids Res 2012;40(15):e115.
[34] Hunter CA, Sanders JKM. The nature of .pi.-.pi. interactions. J Am Chem Soc 1990;112(14):5525–34.
[35] Liu Z, Lu T, Chen Q. Intermolecular interaction characteristics of the all-carboatomic ring, cyclo 18 carbon: focusing on molecular adsorption and stacking. Carbon 2021;171:514–23.
[36] McGaughey GB, Gagné M, Rappé AK. π-stacking interactions: alive and well in proteins. J Biol Chem 1998;273(25):15458–63.
[37] Banerjee A, Anand M, Kalita S, Ganji M. Single-molecule analysis of DNA base-stacking energetics using patterned DNA nanostructures. Nat Nanotechnol 2023:1474–82.
[38] Kool ET. Hydrogen bonding, base stacking, and steric effects in DNA replication. Annu Rev Biophys Biomol Struct 2001;30:1–22.
[39] Zacharias M. Base-pairing and base-stacking contributions to double-stranded DNA formation. J Phys Chem B 2020;124(46):10345–52.
[40] Richardson SM, Wheelan SJ, Yarrington RM, Boeke JD. GeneDesign: rapid, automated design of multikilobase synthetic genes. Genome Res 2006;16(4):550–6.
[41] Bode M, Khor S, Ye H, Li MH, Ying JY. TmPrime: fast, flexible oligonucleotide design software for gene synthesis. Nucleic Acids Res 2009;37(Web Server issue):214–21.
[42] Swainston N, Currin A, Day PJ, Kell DB. GeneGenie: optimized oligomer design for directed evolution. Nucleic Acids Res 2014;42(Web Server issue):395–400.
[43] Plesa C, Sidore AM, Lubock NB, Zhang D, Kosuri S. Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. Science 2018;359(6373):343–7.
[44] Sidore AM, Plesa C, Samson JA, Lubock NB, Kosuri S. DropSynth 2.0: high-fidelity multiplexed gene synthesis in emulsions. Nucleic Acids Res 2020;48(16):e95.
[45] Gorb L, Pekh A, Nyporko A, Ilchenko M, Golius A, Zubatiuk T, Zubatyuk R, Dubey I, Hovorun DM, Leszczynski J. Effect of microenvironment on the geometrical structure of d(A)5 d(T)5 and d(G)5 d(C)5 DNA mini-helixes and the Dickerson dodecamer: a density functional theory study. J Phys Chem B 2020;124(42):9343–53.
[46] Hamlin TA, Poater J, Guerra CF, Bickelhaupt FM. B-DNA model systems in non-terran bio-solvents: implications for structure, stability and replication. Phys Chem Chem Phys 2017;19(26):16969–78.
[47] Nieuwland C, Hamlin TA, Guerra CF, Barone G, Bickelhaupt FM. B-DNA structure and stability: the role of nucleotide composition and order. Chemistryopen 2022;11(2):e202100231.
[48] Baoutina A, Bhat S, Partis L, Emslie KR. Storage stability of solutions of DNA standards. Anal Chem 2019;91(19):12268–74.
[49] Blake RD, Delcourt SG. Thermal stability of DNA. Nucleic Acids Res 1998;26(14):3323–32.
[50] Nisoli C, Bishop AR. Thermomechanics of DNA: theory of thermal stability under load. Phys Rev Lett 2011;107(6):068102.
[51] Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera - a visualization system for exploratory research and analysis. J Comput Chem 2004;25(13):1605–12.
[52] Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015;1–2:19–25.
[53] Ivani I, Dans PD, Noy A, Perez A, Faustino I, Hospital A, Walther J, Andrio P, Goni R, Balaceanu A, Portella G, Battistini F, Lluis Gelpi J, Gonzalez C, Vendruscolo M, Laughton CA, Harris SA, Case DA, Orozco M. Parmbsc1: a refined force field for DNA simulations. Nat Methods 2016;13(1):55–8.
[54] Price DJ, Brooks CL. A modified TIP3P water potential for simulation with Ewald summation. J Chem Phys 2004;121(20):10096–103.
[55] Lu X-J, Olson WK. 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. Nat Protoc 2008;3(7):1213–27.
[56] O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: an open chemical toolbox. J Cheminf 2011;3:33.
[57] Neese F. Software update: the ORCA program system-Version 5.0. Wiley Interdiscip Rev Comput Mol Sci 2022;12(5):e1606.
[58] Chai J-D, Head-Gordon M. Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. Phys Chem Chem Phys 2008;10(44):6615–20.
[59] Weigend F, Ahlrichs R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy. Phys Chem Chem Phys 2005;7(18):3297–305.
[60] Goerigk L, Grimme S. A thorough benchmark of density functional methods for general main group thermochemistry, kinetics, and noncovalent interactions. Phys Chem Chem Phys 2011;13(14):6670–88.
[61] Lu T, Chen F. Multiwfn: a multifunctional wavefunction analyzer. J Comput Chem 2012;33(5):580–92.
[62] Mardirossian N, Head-Gordon M. ωB97X-V: a 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy. Phys Chem Chem Phys 2014;16(21):9904–24.
[63] Mardirossian N, Head-Gordon M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. Mol Phys 2017;115(19):2315–72.
[64] Kossmann S, Neese F. Comparison of two efficient approximate Hartee-Fock approaches. Chem Phys Lett 2009;481(4–6):240–3.
[65] Turney JM, Simmonett AC, Parrish RM, Hohenstein EG, Evangelista FA, Fermann JT, Mintz BJ, Burns LA, Wilke JJ, Abrams ML, Russ NJ, Leininger ML, Janssen CL, Seidl ET, Allen WD, Schaefer HF, King RA, Valeev EF, Sherrill CD, Crawford TD. PSI4: an open-source ab initio electronic structure program. Wiley Interdiscip Rev Comput Mol Sci 2012;2(4):556–65.
[66] Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. J Mol Graph 1996;14(1):33–8. 27-28.
[67] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat I, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy C. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 2020;17(3):261–72.