ChatGPT

Comparative Analysis: DNA-Diffusion and DiscDiff in Genomic Sequence Generation

Introduction and Overview

Generative AI models are increasingly being applied to **genomic sequences** for tasks like designing synthetic DNA with desired regulatory functions. Two recent approaches – **DNA-Diffusion** and **DiscDiff** – explore *diffusion models* for DNA sequence generation from different angles. **DNA-Diffusion** (2024) leverages a diffusion model to **generate synthetic regulatory DNA elements** (short sequences ~200 bp) conditioned on cell type, aiming to control chromatin accessibility and gene expression [1]. **DiscDiff** (2024) introduces a framework for **DNA sequence generation across species** using a latent diffusion model, coupled with a post-processing algorithm, to produce realistic DNA sequences (including regulatory regions and gene sequences) with high fidelity [2]. Both models set new milestones in DNA generative modeling: DNA-Diffusion demonstrates the ability to create cell type-specific enhancers in silico [3], while DiscDiff achieves state-of-the-art performance in generating both short and long DNA sequences across multiple species [2]. Despite sharing the diffusion paradigm, they differ in methodology and objectives – DNA-Diffusion focuses on **conditional generation** of regulatory sequences for synthetic biology, whereas DiscDiff proposes a **latent discrete diffusion** approach to broadly model genomic sequences. Below, we discuss each model's achievements and methods, compare their similarities and differences (including how DiscDiff tackles a key concern in DNA-Diffusion's design), and examine why diffusion models were chosen over traditional sequence models. We also highlight the data used (clarifying terms like *"endogenous"* data), connections to prior DNA language models (e.g. **Evo 2**), and the limitations and potential applications of these generative genomics models in the context of *AI for virtual cells* and beyond.

DNA-Diffusion: Diffusion for Cell Type-Specific Regulatory DNA Design

DNA-Diffusion (Ferreira DaSilva *et al.* 2024) is a conditional diffusion model developed to **design synthetic DNA regulatory elements (enhancers)** that can drive gene expression in a **cell type-specific** manner [1]. The model's goal is to aid synthetic biology and gene therapy by generating 200 bp DNA sequences with desired regulatory activity – for example, creating an enhancer sequence that is active (open chromatin and transcriptionally potent) in one cell type but not in others [1] [4]. To achieve this, DNA-Diffusion uses the standard diffusion model framework adapted to one-dimensional DNA sequences:

- **Architecture & Method:** DNA-Diffusion employs a **U-Net convolutional network** (inspired by image diffusion models) that iteratively denoises a noisy sequence input [5] [6]. Each DNA sequence is represented as a one-hot encoded matrix (4 channels for A/C/G/T by 200 bp length). During the *forward diffusion* process, **Gaussian noise** is added to these one-hot sequence vectors over multiple timesteps until the sequence becomes nearly random noise [7]. Training involves learning the *reverse diffusion* – the U-Net is given a noisy sequence plus a time-step embedding and a **cell type condition** label, and it learns to predict and remove the noise to recover the original

sequence at that timestep [8] [9]. Essentially, the model is **conditioned on cell type** so that it will generate sequences with features specific to that cell type's open chromatin profiles. After training, new sequences can be generated by starting from random noise and iteratively denoising it *50 steps* down to a synthetic DNA sequence that "reflects the characteristics of the target cell type" [9]. The diffusion model "receives three inputs: DNA sequences, a timestep, and cell type labels" and uses the cell-type embedding to guide the sequence generation toward the desired regulatory profile [7].

- **Key Achievement:** DNA-Diffusion showed that diffusion models can **"robustly generate DNA sequences with cell type-specific regulatory potential."** In evaluations, the authors verified that the synthetic sequences retained hallmark properties of real (*endogenous*) regulatory DNA. For instance, generated sequences had realistic combinations of transcription factor binding sites and were predicted (by external models) to produce the intended chromatin accessibility and gene expression patterns in the target cell types [10]. In fact, using state-of-the-art predictive models (like Enformer/AlphaGenome-style epigenomic predictors), they found that DNA-Diffusion sequences could **activate genes and open chromatin** in silico similarly to or even beyond real enhancers for those cell contexts [11] [12]. This demonstrates the potential of the model to *modulate gene expression* via designed DNA, paving the way for applications in mammalian synthetic biology and precision gene therapy [4].

- **Data Used (Endogenous vs Synthetic):** DNA-Diffusion was trained on **real genomic sequences** – specifically, **DNase hypersensitive sites (DHS)** of length ~200 bp from a few well-studied human cell lines (GM12878 lymphoblastoid cells, K562 erythroleukemia cells, and HepG2 liver cells) [13]. These **"endogenous" sequences** are actual regulatory DNA elements native to the human genome that were accessible (open chromatin) in specific cell types, drawn from a DHS index dataset (ENCODE/Roadmap epigenomics data) curated by Meuleman *et al.* [13]. In the context of the paper, *"endogenous" data refers to real biological sequences present in the genome*, as opposed to *synthetic sequences* generated by the model. The authors ensured that the diffusion model's outputs were biologically plausible by comparing them to such endogenous sequences. Notably, they checked that the **synthetic enhancers** did not trivially copy training examples – out of hundreds of thousands of generated sequences, only a handful were exact matches to known DHS sequences (e.g. ~50 overlaps for GM12878, indicating high novelty/diversity) [14] [15]. This indicates the model was creating novel sequence variants rather than memorizing, a critical property for generative design.

- **Addressing Diffusion on Discrete Data:** One methodological detail is that DNA-Diffusion applied **Gaussian noise to one-hot vectors** – meaning during noising, a DNA sequence (which initially has binary 0/1 values in each nucleotide channel) becomes a continuous valued matrix [7]. This is a straightforward adaptation of image/text diffusion to DNA, but it introduces *intermediate states that are not valid DNA sequences* (e.g. partial activation of multiple nucleotide channels). The model must learn to navigate these continuous states and ultimately produce a valid discrete sequence (usually by taking the argmax base at each position after the final denoising step). The authors used a fixed noise schedule (50 diffusion steps) and trained the U-Net to predict the added noise at each step so that it can subtract it out [8]. While effective, **using standard normal noise on one-hot encodings** can cause a kind of *"rounding" problem*: the model's output before discretization might be ambiguous or unrealistic (like 0.5 A + 0.5 C at one position). DNA-Diffusion dealt with this by relying on the neural network to produce sharp outputs that can be thresholded into A/C/G/T. However, this approach left a space for improvement – as we will see, DiscDiff explicitly tackles the challenge of

discrete sequence diffusion by altering how the diffusion is performed (hint: via a latent space encoding).

In summary, DNA-Diffusion's methodology demonstrated **how diffusion models can be adapted to DNA sequence generation** with conditioning. It achieved the generation of functional, cell-specific DNA sequences and established a baseline for *controlling gene regulatory elements via generative models*. One primary concern with this approach is the continuous noising of discrete data, which can introduce inaccuracies. The **DiscDiff** paper addresses this concern through an alternative diffusion strategy.

DiscDiff: Latent Diffusion Model for DNA Sequence Generation

DiscDiff (Li *et al.* 2024) presents a novel two-part framework for DNA generation, consisting of **(1) DiscDiff**, a latent diffusion model (LDM) tailored to discrete sequences, and **(2) Absorb-Escape**, a post-processing algorithm to refine generated sequences [2]. The motivation behind DiscDiff is to overcome the unique challenges of applying diffusion models to DNA, particularly the *discrete nature* of nucleotide sequences and the need for diverse, high-fidelity outputs across different sequence types (regulatory regions and protein-coding sequences, potentially in multiple species). DiscDiff's approach can be seen as a more *general-purpose DNA generator*: it is trained on a large multi-species DNA dataset and can generate both short sequences (like promoters/enhancers) and longer sequences (gene regions up to 2 kb) [16] [17]. Importantly, DiscDiff directly addresses the **"continuous vs discrete" mismatch** present in models like DNA-Diffusion by introducing a latent representation for sequences and a correction mechanism for any errors in conversion between latent and actual DNA.

- **Latent Diffusion Methodology:** Rather than diffusing on one-hot DNA directly, DiscDiff first **encodes DNA sequences into a continuous latent space** using a specialized autoencoder (in fact, they experiment with architectures including CNNs and transformers – e.g. a Swin-transformer – for the encoder/decoder, even noting that incorporating components like those in Enformer improved performance in one variant) [18]. The encoder compresses a DNA sequence into a latent vector (of lower dimension), and the decoder can reconstruct the sequence from this latent (this is akin to a VAE, but they focus on minimizing reconstruction error). Once this **DNA-to-latent mapping** is established, a **diffusion model is trained in the latent space**: Gaussian noise is added to the latent vectors rather than to the raw sequence, and a diffusion U-Net (or similar) learns to denoise in latent space [19] [20]. Because the latent space is continuous by construction, the diffusion process here is naturally handled with continuous noise, avoiding the problem of generating physically impossible intermediate DNA states. After training, DiscDiff generates new sequences by sampling a random latent vector, iteratively denoising it via the diffusion model, and then passing the final denoised latent through the decoder to produce a DNA sequence.

- **Absorb-Escape (Refinement Step):** A critical innovation in DiscDiff is the **Absorb-Escape algorithm**, which aims to fix any *"rounding errors"* or local inconsistencies that arise when converting the denoised latent back into a discrete sequence [2] [21]. Even with a good autoencoder, the decoded sequence might have minor errors (e.g. low-probability nucleotides where there should be a clear choice). Absorb-Escape addresses this by leveraging an **autoregressive model** in a post-hoc manner. Specifically, they use **Hyena**, a state-of-the-art autoregressive sequence model, to scan through the generated sequence and correct mistakes. The Absorb-Escape algorithm works by *"absorbing"* the reliable parts of the sequence and *"escaping"* (redoing) the uncertain parts, effectively combining the **strengths of diffusion (global realism and diversity)** with the **strengths of autoregressive**

models (local coherence and grammar) [22]. This hybrid approach was shown to significantly improve the quality of generated sequences – for example, Absorb-Escape improved DiscDiff's performance by ~4% on long sequences and enhanced the model's ability to precisely match known motif patterns [23] [24]. In essence, DiscDiff generates a draft sequence in one shot via latent diffusion, and then Absorb-Escape fine-tunes that sequence to eliminate small errors, much like an editor refining a rough draft.

- **Achievements and Evaluation:** DiscDiff is reported to **outperform prior generative models** for DNA in both **short sequence generation** (enhancers, promoters) and **long sequence generation** (gene regions) [25]. It was evaluated on a suite of metrics, including a *latent Frechet distance* (S-FID) measuring how close the distribution of generated sequences is to real sequences in a learned feature space, and **motif distribution correlation**, which checks if the frequency of biological motifs (like transcription factor binding sequences) in generated DNA matches those in real DNA [26]. DiscDiff had the **smallest latent distance and highest motif correlation** compared to other baseline models (including earlier diffusion models and an autoregressive transformer), indicating its samples are both realistic and capture the underlying genomic patterns very well [26]. Notably, the authors compared DiscDiff to a discrete diffusion model (DDSM) and to a transformer-based generative model, and DiscDiff achieved the best trade-off of accuracy and diversity [27] [28]. They also performed a head-to-head comparison with an autoregressive DNA language model (Hyena) on conditional generation tasks; each had strengths for certain motif types, but combining them via Absorb-Escape yielded the most realistic sequences overall [29] [30].

- **Data Used:** A major contribution of DiscDiff is the introduction of **EPD-GenDNA**, a comprehensive multi-species dataset of DNA sequences for generative modeling [31] [32]. EPD-GenDNA is built from the Eukaryotic Promoter Database and contains **160,000 unique sequences** from **15 species**, including both **regulatory regions (promoters/enhancers)** and **protein-coding segments** [33]. Sequence lengths in the dataset are either 256 bp (promoter-centered windows) or 2048 bp (extended genomic regions around transcription start sites) [34]. Each sequence comes with rich annotations like species and cell type of origin and gene expression info [35] [36]. This dataset allowed DiscDiff to be the **first DNA diffusion model tested across multiple species** (prior works were largely single-species and much smaller) [17]. By training on this diverse data, DiscDiff learned to generate DNA that is not just human-like but captures broader genomic "languages." The diversity of training data is reflected in its outputs: DiscDiff can sample novel DNA sequences that maintain **natural diversity** (measured by unique *n*-gram content close to real data) [37]. In fact, the authors emphasize that **diversity is a key metric for genome sequence generation**, to avoid mode collapse and to ensure synthetic sequences explore the full space of possibilities [38]. DiscDiff's diffusion approach, combined with latent compression, was explicitly designed to promote diversity while still matching biological distributions.

- **Key Differences from DNA-Diffusion:** DiscDiff directly tackled the main concern one might have with a method like DNA-Diffusion: *the use of standard Gaussian noise on discrete sequences*. In DNA-Diffusion, adding continuous noise to one-hot DNA could lead to off-manifold intermediate states that the model must correct. DiscDiff avoids diffusing on raw sequences; instead, it **performs diffusion in a learned continuous latent space** where noise application is natural [39]. The consequence is fewer downstream errors when decoding sequences. And if errors do occur, the Absorb-Escape step fixes them, resulting in highly realistic final sequences [21]. In summary, DiscDiff's methodology is more complex (involving an encoder, a diffusion model, and an AR refiner),

but this complexity is justified by significantly improved quality and the ability to handle long sequences. DiscDiff also operates largely *unconditionally* or with broad conditioning (it can generate sequences without needing a specific cell type label, or potentially conditioned on species or desired motifs), whereas DNA-Diffusion was a fully **conditional model (cell type-specific)**. Despite these differences, both models share the core idea that **diffusion models can generate diverse, high-fidelity DNA** – they just implement it in different ways to accommodate the discrete sequence data.

Why Diffusion Models? (Diffusion vs. Autoregressive Transformers)

A central question is why these works chose *diffusion models* for DNA generation instead of more traditional autoregressive (AR) sequence models like Transformers or RNN-based language models. Several reasons emerge from the papers:

- **Avoiding Mode Collapse & Enhancing Diversity:** Earlier attempts at DNA generation used Generative Adversarial Networks (GANs) or even direct CNN-based generators, but those often suffered from *mode collapse* and limited diversity – i.e. they tended to produce very similar or repetitive sequences, failing to cover the full variability of genomic patterns [40]. Autoregressive language models can also become over-confident and repetitive, essentially learning to generate the most frequent patterns and ignoring rarer variants [41]. In genomic sequence generation, such loss of diversity is *"detrimental"* because capturing the wide range of possible sequences (especially for non-coding DNA where many different motifs and combinations exist) is crucial [41]. For instance, human individuals share ~98% of their genome, but the 2% that varies encompasses millions of distinct variants; a generative model should be able to produce many of these potential variations, not just regurgitate the reference sequence. **Diffusion models naturally encourage diversity** by their stochastic sampling process – each diffusion run can produce a different outcome, and there's no easy way for the model to collapse to a single mode since it has to learn to reconstruct data from random noise. Dhariwal & Nichol (2021) observed that diffusion models can outperform GANs in sample diversity [40]. In the context of DNA, DiscDiff explicitly highlights that autoregressive transformers *"generate samples with lower diversity than diffusion models"* and tend to repeat themselves, whereas diffusion models produce a richer variety of outputs [41]. This is a major motivation: **to generate truly novel DNA sequences (for synthetic biology or data augmentation), one needs the generative model to explore many modes of the sequence distribution**, which diffusion is well-suited for.

- **Iterative Refinement and Global Coherence:** Diffusion models generate data by iterative denoising, which allows them to make large-scale decisions in a sequence in a coarse-to-fine manner. This can help with **global coherence** of the sequence in a way that one-base-at-a-time generation sometimes struggles with. An AR model decides each next nucleotide based only on previous ones; if it makes a bad choice early, it can derail the rest of the sequence (accumulating errors) [42]. This can lead to issues like local repetitions or inconsistent long-range structure in AR outputs [38]. Diffusion, in contrast, starts with a holistic (if noisy) view and refines everything together, potentially balancing local and global features as it denoises. In DNA-Diffusion's case, the U-Net has access to the entire sequence and the cell type context at once, so it can, for example, ensure that if a certain transcription factor motif is needed at the start and another at the end, it can place both during the denoising process. This iterative global refinement tends to produce sequences that capture higher-order dependencies (like spacing between motifs, or overall GC-content) more naturally. That said, diffusion models sometimes *under-perform in local syntax* (e.g.

occasional minor errors in a sequence) compared to AR models, which is why DiscDiff introduced Absorb-Escape to inject local autoregressive strength into the final output [22].

- **Handling Long Contexts:** Genomic sequences can be very long (thousands to millions of bases). Autoregressive transformers face **computational challenges** with long contexts due to quadratic attention scaling and memory limits. Notably, the Evo 2 language model (discussed below) addresses this with 1 million token context, but it requires enormous model sizes (billions of parameters) and specialized training regimes [43] [44]. Diffusion models, especially in latent space, offer an alternative: compress the sequence and then generate. DiscDiff's latent diffusion is an example – by encoding 2048bp sequences into a latent, they sidestep dealing with thousands of tokens in the generative model directly [19]. This makes the generation of long sequences more tractable on modest computational budgets. In effect, diffusion models can leverage **latent representations and iterative sampling** to cover long-range genomic structure without needing an exorbitantly large model at generation time. This is partly why DiscDiff could demonstrate multi-kilobase sequence generation with high fidelity [25].

- **Conditional Generation is Straightforward:** Both DNA-Diffusion and DiscDiff needed to condition on certain information (cell type for DNA-Diffusion; in DiscDiff's case, they considered conditional generation in experiments, such as giving the model a motif to include or a species label). In diffusion models, conditioning can often be done by concatenating the condition to the input or modifying the denoising network (e.g. through cross-attention or FiLM layers) without fundamentally changing the generation process. DNA-Diffusion simply concatenated a one-hot cell type vector as additional input channels to the U-Net and as a global label, which is a natural extension of image diffusion techniques [8]. In contrast, conditioning an autoregressive DNA model on metadata might require special tokens or complex prompt design, and ensuring the model actually respects the condition can be tricky. The diffusion approach thus provided a clean way to inject context like "generate an enhancer for K562 cells" into the process. Additionally, diffusion models can integrate multiple conditions (if needed) by altering the denoising score function, offering flexibility for future extensions (e.g. generate a sequence with **both** a certain GC content **and** a specific histone mark profile – one could condition the diffusion model on both).

- **Discrete Data Challenges and Solutions:** A noteworthy point is that *vanilla* diffusion was designed for continuous data (like pixel intensities). Applying it to DNA (discrete alphabet) is non-trivial, and yet both papers show viable solutions. DNA-Diffusion took the simpler route of using continuous noise on one-hot vectors, effectively *pretending DNA is like an image* and relying on the network to snap it back to one-hot [7] [9]. This works but could introduce errors – imagine the model outputs a 0.8 for "A" and 0.7 for "C" at one position, which base do you choose? It requires a hard decision that could slightly perturb the sequence's validity (this could potentially create an out-of-distribution sequence if not careful). **DiscDiff's latent approach is a more principled solution**: by learning a continuous embedding of DNA, the diffusion model always operates in a continuous space that (ideally) corresponds to valid sequences when decoded. The **problem of "rounding"** (continuous outputs not mapping cleanly to discrete tokens) is alleviated by design and explicitly corrected with Absorb-Escape [21]. Moreover, other research has explored *discrete diffusion* (e.g. **D3PM** or **BitDiffusion** for sequences, and **DDSM** using a Dirichlet distribution for DNA) [45]. These confirm that special noise distributions can be used for discrete data, but they tend to be more complex or less efficient. DiscDiff's contribution was showing that **Latent Diffusion Models (LDMs)** – which were very successful in imaging – can be **adapted to DNA** with new encoder/decoder architectures

and a bit of autoregressive help [46] [47]. This substantially reduces the computational cost compared to operating in the original sequence space (DDSM and related discrete DMs were found to be computationally intensive [48]), making diffusion a practical choice for genomics.

In short, diffusion models were chosen because they offer **greater sample diversity**, **robust global generation** and flexible conditioning, which are valuable for *de novo* DNA design. The tradeoff is dealing with discrete data issues, but as DiscDiff demonstrated, this can be overcome with a creative combination of techniques. The result is generative models that can imagine a wide variety of realistic DNA sequences – something that a naïve transformer might struggle with, either collapsing to repetitive motifs or requiring an extremely large model to capture all variations [49]. It's worth noting that **the DiscDiff authors still acknowledge the strengths of autoregressive models**, which is why they integrated Hyena to handle fine details [22]. Thus, the emerging view is that **diffusion and autoregressive approaches can be complementary** for biological sequence generation [50]. Some recent works even try to combine them (e.g. using a language model to guide diffusion) [50], and DiscDiff's Absorb-Escape is a novel instance of such a combination.

Data Sources and "Endogenous" Sequence Properties

Both models rely on high-quality genomic data and careful evaluation to ensure their outputs are biologically meaningful:

- **DNA-Diffusion data:** as mentioned, it uses **endogenous DHS sequences** from real cells. These are sequences which were experimentally identified as open chromatin (potential enhancers or promoters) in specific cell types [13]. By training on these, the model learns the "signature" sequence features of regulatory DNA for each cell type (for example, K562 cells might enrich GATA1 motifs, etc.). *Endogenous* here simply means these sequences come from the organism's own genome (the term is used to contrast with *exogenous* synthetic sequences). The authors evaluated whether generated sequences kept *"properties of endogenous sequences"* – meaning do they look and act like real DNA elements? They examined things like **transcription factor binding site composition** (does the sequence have the right kinds and number of motifs that real enhancers have?), **predicted chromatin accessibility** (if you feed the sequence into a predictive model for chromatin or ATAC-seq, does it score as open in the target cell and closed in others?), and **predicted gene expression activation** (using e.g. a sequence-to-gene expression model to see if the sequence can drive expression of a reporter) [10]. These metrics were computed with *state-of-the-art predictive models* – for instance, they used **Enformer/DeepSpeed (Avsec et al. 2021)** or similar deep learning models trained to predict epigenomic signals from sequence, effectively acting as an *in silico assay* for the synthetic DNA. By these evaluations, DNA-Diffusion sequences performed impressively: they often matched the cell-type specificity of real enhancers and in some cases were able to **induce stronger predicted activity** than actual genomic sequences when placed in certain genomic contexts [11] [12]. This suggests the model was not just memorizing sequences, but capturing general rules (e.g. motif grammar) to invent new possibly "better" enhancers. It underscores the **application potential**: one could design enhancers that are optimized for a purpose (like maximal gene activation in cell type X) which might not exist exactly in nature.

- **DiscDiff data:** EPD-GenDNA, being cross-species, gives DiscDiff a very broad training distribution – from human to fruit fly to yeast DNA [16]. The sequences include **promoter regions near transcription start sites (TSS)**, which encompass both regulatory elements and the beginning of

gene coding sequences [34]. Each sequence in the dataset is annotated with the species and sometimes the cell type or tissue in which the promoter was characterized [35]. This allowed the authors to test both **unconditional generation** (generate a random DNA sequence from the learned genomic distribution) and **conditional generation** (generate a sequence for a given species or with a given property) [51] [52]. For example, they could ask DiscDiff to generate 256bp sequences that resemble *Drosophila* promoters versus human promoters and assess differences. They quantitatively evaluated DiscDiff's outputs against real data for each species, using metrics like k-mer diversity and motif occurrence, as mentioned earlier. They found DiscDiff could closely replicate the *species-specific* characteristics of sequences – e.g. it captured that yeast promoters have different sequence content than mammalian promoters – yet still produce novel sequences that weren't in the training set [53] [54]. This speaks to the model's ability to generalize. The dataset's size (160k sequences and augmented to 30 million samples through data augmentation techniques [34]) provided enough complexity for training a diffusion model without severe overfitting, and indeed DiscDiff's diversity metric was very close to natural diversity (only a small divergence, and better than other models) [37] [55].

In both cases, having a **large and representative dataset** was crucial. DNA-Diffusion's reliance on ENCODE/Roadmap data ensured the model learned real genomic "grammar." DiscDiff's creation of a new dataset filled a gap – previously, generative DNA studies had to make do with small, single-species datasets (like a few thousand sequences of yeast or human enhancers) [56]. By providing EPD-GenDNA, they established a benchmark for future models and ensured DiscDiff had enough training data to be effective [57] [58]. It's also worth clarifying that DiscDiff's dataset includes *both regulatory and protein-coding segments*, meaning the model isn't limited to enhancers; it also sees gene exon sequences. This might allow it to generate plausible coding sequences (for protein production applications) as well as regulatory DNA [17]. The inclusion of multiple species is advantageous because it forces the model to capture more fundamental patterns (conserved signals like TATA-box, etc., that appear across species, as well as species-specific biases). In essence, the data breadth acts as a form of regularization and enrichment for the generative model.

Finally, the notion of *"endogenous"* sequences in DiscDiff's context would refer to the real promoter sequences from EPD. The authors did compare generated sequences to these real ones to ensure things like **motif distributions** line up [26]. The generative model should ideally produce sequences indistinguishable (to an expert or a classifier) from actual genomic sequences. DiscDiff largely achieved this, with the help of Absorb-Escape to correct any tell-tale errors. The success of both models in matching endogenous sequence properties gives confidence that these synthetic sequences could function biologically – at least as far as *in silico* predictions indicate.

Comparison with Prior and Related Models (Evo 2, Autoregressive Models, etc.)

Both DNA-Diffusion and DiscDiff build upon and differ from previous approaches to modeling DNA sequences:

- **Versus Autoregressive "DNA Language Models":** Prior to diffusion models, many works applied language-model style architectures to DNA. For example, recurrent networks or Transformers have been used to learn genome sequences and even generate them or score them (e.g. *Genomic GPTs*,

DNABERT for classification, or generative LSTMs for DNA). As DiscDiff's related work notes, large autoregressive models (including Transformers and state-space models) have indeed been trained on genomic sequences [59]. They can capture sequence dependencies and have seen success in tasks like predicting functional genomic signals or generating DNA one base at a time. **Evo 2** is a very recent example of a high-profile AR model: *Evo 2* (Brixi *et al.* 2025) is described as a **"biological foundation model"** trained on an unprecedented **9.3 trillion DNA bases** across all domains of life [43]. It's essentially a gargantuan DNA language model with up to **40 billion parameters** and a context window of **1 million bp** [43] [44]. Evo 2 is designed to both predict functional effects of genetic variation and *generate genomic sequences*, covering bacteria, archaea, and eukaryotes in one model [43]. In concept, Evo 2 and DiscDiff share a grand vision: **modeling genomes across all life for design and prediction**. However, their approaches differ radically:

- *Evo 2* uses an **autoregressive transformer** (and possibly specialized long-context mechanisms) to directly model DNA sequences as text, requiring enormous data and compute to achieve its performance [43].
- *DiscDiff* uses a **latent diffusion** with a comparatively compact model, focusing on a curated set of eukaryotic promoters (160k sequences is tiny compared to 9.3e12 bases) [33]. It relies on clever architecture (autoencoder + diffusion) rather than brute-force scale. As a result, DiscDiff can't handle million-base contexts explicitly – it's limited to the sequence lengths it was trained on (up to 2 kb) – whereas Evo 2 can in principle model whole genomic loci with length up to 1 Mb or more [43].
- **Similarities:** Both aim to **generate realistic genomic sequences** and can be used for **genome design** tasks. They both incorporate **multi-species data**: Evo 2 spans *all* domains of life, DiscDiff spans 15 species including multiple vertebrates and model organisms [33]. This multi-species training gives both models a sense of evolutionary diversity and constraint. Another similarity is that **Evo 2 and DiscDiff can predict functional information** about sequences: Evo 2's foundation model can be queried for variant effects (the paper reports state-of-the-art performance on many variant effect prediction benchmarks) and can generate sequences with certain properties, while DiscDiff + Absorb-Escape demonstrated the ability to conditionally generate sequences with desired motif patterns or from certain species, and they mention potential use in gene therapy (implying designing sequences with function in mind) [60].
- **Differences:** The methodological difference (AR vs diffusion) means Evo 2 may face the issues of AR models (it might need measures to maintain diversity and avoid collapse). However, given its scale, Evo 2 likely mitigates some of this by sheer capacity – it can memorize and reproduce a vast array of patterns. Still, DiscDiff's results suggest that a far smaller diffusion model was able to achieve very high fidelity on a focused task [25]. Another difference is interpretability: a latent diffusion model is a bit of a black box in how it composes sequence features, whereas a transformer model's attention might be interrogated to see what motifs or dependencies it's attending to (though at 40B params that's also difficult). **Training data** is a huge differentiator: Evo 2 was trained on essentially all available genomes (billions of sequences), while DiscDiff trained on a specific promoter database. Therefore, Evo 2 might capture things like non-coding repeat structure, large-scale genome organization, etc., which DiscDiff would not see. On the flip side, DiscDiff's training on manually curated promoters means it learned very high-quality, functional sequences; Evo 2, by casting such a wide net, might have to filter out or down-weight non-functional or non-genic DNA.

- **How they can be related:** We can envision using these models in tandem or comparing insights. For example, DiscDiff could generate candidate promoter sequences and Evo 2 (or another foundation model) could evaluate their likelihood or functional scores (since Evo 2 can predict epigenomic signals, similar to AlphaGenome described below). Conversely, one could use Evo 2 to suggest a rough sequence design and then use DiscDiff's refinement (via diffusion + Absorb-Escape) to polish

certain regions. Both represent steps toward **AI-driven genome design**. It's notable that Evo 2's authors explicitly mention enabling "genomic and epigenomic design" as a goal [43] – a goal very much in line with what DNA-Diffusion and DiscDiff are doing on a smaller scale. In summary, DiscDiff and Evo 2 share the vision but differ in strategy: *DiscDiff demonstrates that with the right model design, even relatively small datasets can yield powerful generative DNA models*, whereas *Evo 2 demonstrates the power of scale and breadth, treating genome modeling as a language modeling problem at internet-scale data*. Time and research will tell which approach or combination yields the most utility for scientists.

- **Versus Other Diffusion Models (and prior generative work):** DiscDiff wasn't the first attempt to adapt diffusion to biological sequences. The authors cite **DDSM (Avdeyev et al. 2023)**, a Dirichlet Diffusion Score Model for human DNA [45]. DDSM used a *Dirichlet noise distribution* to diffuse probability vectors over the DNA alphabet, thereby maintaining a valid distribution at each step (no need for Gaussian with one-hots). It achieved good results especially on conditional tasks, but DDSM still operated in the original sequence length (1024 bp human sequences) and was computationally heavy [48]. DiscDiff differentiates itself by using latent compression (which DDSM did not) and by demonstrating use across species and both coding/regulatory regions [17]. Additionally, **EvoDiff (Alamdari et al. 2023)** is mentioned – EvoDiff was a diffusion model for **protein sequences** (amino acid sequences) which also had to handle discrete tokens [48]. EvoDiff and related models (like ProteinMPNN or Score-based generative models for proteins) illustrate that diffusion can explore sequence space in biology broadly. DiscDiff takes inspiration from these but tailors the approach to DNA. It introduces domain-specific solutions like Absorb-Escape, which were not present in EvoDiff. In terms of performance, DiscDiff reported superior results to previous diffusion models in DNA (including DDSM and a BitDiffusion adaptation) [61]. It also compared against *D3PM* (a discrete diffusion framework from Austin et al. 2021) and found DiscDiff's outputs had closer motif statistics to the ground truth [37] [55].

- **Versus GANs and heuristic sequence design:** Before diffusion models, GANs were tried for DNA – e.g. **EnhancerGAN, FeedbackGAN** for enhancer or promoter generation [62]. These had some success but often produced sequences that, upon evaluation, lacked the full diversity of real data or contained artifacts. For instance, one might get a few strong motifs repeated too often (mode collapse) or sequences that did not integrate multiple regulatory signals well. Both DNA-Diffusion and DiscDiff make the case that diffusion models overcome these issues, producing more varied and realistic sequences. The quantitative metrics in DiscDiff (like diversity and S-FID) back this claim, as DiscDiff's diversity was very close to the real dataset whereas GAN baselines were presumably worse (the paper notes GAN outputs were less diverse and had mode collapse issues [40]). Another advantage is training stability: GANs are notoriously fickle to train, whereas diffusion model training is a straightforward maximum likelihood training (denoising score matching) which is generally stable albeit slow. This reliability is important if we want a method that can scale or be used as a basis in many labs.

In summary, **DiscDiff stands out as the first latent diffusion for DNA,** and it set a new bar by combining best practices from prior art (latent modeling, discrete diffusion strategies, AR refinement) and contributing new datasets and algorithms. **DNA-Diffusion** stands out as one of the first to *explicitly tackle controlled regulatory sequence generation*, demonstrating a clear use-case (synthetic enhancers) and validating it with downstream functional predictions [3]. They both represent a departure from purely autoregressive or GAN approaches, harnessing diffusion to great effect. The emergence of **Evo 2** (and also DeepMind's **AlphaMissense** or **AlphaGenome** for predictions) indicates that huge AR models are another path to

genome modeling. It will be fascinating to see integration: for example, could a future model use Evo 2's knowledge to guide a diffusion model? Or use diffusion models to fine-tune large language model outputs? The field is moving quickly, and these works collectively inspire hybrid approaches (indeed, DiscDiff is itself a hybrid of diffusion + AR).

Limitations and Applications – "So What" of Generative Genomics?

While these diffusion models for DNA are technically impressive, one might ask: **what are the practical applications of generating DNA sequences, and what limitations remain?** The papers provide some hints, and we can extrapolate possible uses in research and biotechnology:

- **Synthetic Biology & Gene Therapy Design:** A primary motivation for DNA-Diffusion was to *enable precise control of gene expression* through synthetic DNA. In practice, this could mean designing an **enhancer or promoter for a therapeutic gene** – for example, creating a regulatory sequence that only turns on a gene in T-cells but not in other cells, or an enhancer that is active only in low-oxygen conditions, etc. Previously, one would have to screen many candidate sequences or rely on minimal motif tweaking. DNA-Diffusion provides a way to **algorithmically design such sequences** by specifying the desired context (cell type) and letting the model generate candidates. This is useful for building gene therapies that need cell-specific targeting or for **engineering cell-based therapies** where you want a gene to be tightly controlled. Another application is in **biomanufacturing**: designing promoters that maximize production of a protein in a cell line (the model could be conditioned on a context representing high expression, for instance). DiscDiff's authors explicitly mention *"potential implications for gene therapy and protein production"* [60] – imagining that their model (which can generate entire promoters or even genes) could be used to create novel gene constructs optimized for these purposes. For protein production, one might use DiscDiff to generate coding sequences (synonymous variants of a gene) that have favorable codon usage or mRNA structure for high translation efficiency. Because DiscDiff saw many coding regions, it could conceivably generate a gene sequence that is different from any natural gene but still yields a functional protein, perhaps optimized for expression in a certain organism.

- **Data Augmentation and Privacy:** In human genomics research, a big challenge is data sharing and scarcity of labeled examples (and privacy concerns with real genomes). Generative models can create **synthetic genomic data** that mimics real data distribution without containing exact private information. For instance, the **Kenneweg et al. (2025)** work (arXiv 2412.03278, referenced in DiscDiff context) used diffusion to generate **entire synthetic human genotypes** for exactly this purpose – enabling researchers to train models on synthetic genomes when real ones are protected [63] [64]. While DiscDiff itself didn't explicitly demonstrate genome-scale generation, its approach to modeling long sequences and multi-species data is a step in that direction. One can imagine using DiscDiff or similar to generate *synthetic patient genomes* that preserve allele frequency spectra and linkage disequilibrium patterns, which would be extremely valuable for genomic studies while respecting privacy [65] [66]. Even at the smaller scale, synthetic enhancers from DNA-Diffusion could augment datasets for training predictive models: e.g., to train a classifier to identify cell-type-specific enhancers, one could add some model-generated examples to balance classes or explore feature space. Both papers noted that augmenting real data with synthetic data improved performance of downstream models [67] [68]. Thus, generative DNA models can serve as *"data generators"* to bolster learning tasks in genomics.

- **Understanding Regulatory Grammar:** Generative models can be tools for science discovery. By analyzing what the model generates, researchers can infer what patterns it "thinks" are necessary for function. DNA-Diffusion, for example, could be queried to produce multiple enhancers for the same cell type and then one could look for common motifs – if the model consistently inserts a particular motif, it reinforces the evidence that motif is important for that cell type. Conversely, one could ask the model to generate a sequence with certain constraints (if the model supports that) to test hypotheses (e.g., "generate a K562 enhancer sequence that does *not* contain GATA1 motif" – if the model struggles, that suggests GATA1 is essential for K562 enhancers). DiscDiff's Absorb-Escape uses an AR model that could potentially assign probabilities to sections of sequence; low confidence regions might correspond to biologically constrained positions (like the TATA box which must be a specific sequence). In this way, generative models can help **identify key sequence features** and perhaps design experiments. The DNA-Diffusion study itself went a step further: they took synthetic sequences and *experimentally tested them in silico* by inserting them into genomic contexts and using predictive models to see if they activate genes [69] [70]. They even identified cases where synthetic sequences activated genes more strongly than natural ones [71] [72], suggesting we could design "super-enhancers" or novel regulatory switches that nature hasn't utilized. This opens a *creative* side to genomics – using AI to **invent new biological components**.

- **Limitations – Functional Validity:** Despite promising in silico results, a clear limitation is that **experimental validation is needed**. For DNA-Diffusion's sequences, the real test is to synthesize them and put them in cells to see if they indeed produce the chromatin and expression changes predicted. Models like Enformer (and AlphaGenome) are powerful predictors, but they are not perfect; a sequence predicted to be a strong enhancer might not work in vivo due to chromatin context or 3D genome interactions not accounted for. Thus, one limitation is that generative models may produce sequences that look good to current predictive models (which are trained on known biology), but biology could still surprise us with unintended effects (e.g. the sequence might form a secondary structure or be toxic in some way). **DiscDiff's sequences** similarly have not been experimentally tested – they are evaluated by statistical metrics. So a limitation is we don't yet know if DiscDiff can generate, say, a *functional gene* that expresses properly, or a promoter that actually drives transcription in a living cell. It's one thing to match k-mer distributions; it's another to have all subtle features needed for function. Future work likely will involve moving these models "from silico to vivo," testing a sample of generated sequences in wet-lab assays (e.g., massively parallel reporter assays for enhancers) to validate their functionality. Until then, there's a caution that these sequences are *hypotheses* rather than proven designs.

- **Limitations – Sequence Length and Context:** DNA-Diffusion only generates 200 bp sequences in isolation. In reality, an enhancer's effect might depend on surrounding genomic context or synergy with promoters. Likewise, a 2 kb sequence from DiscDiff might represent a promoter plus some upstream sequence, but it's still not a full genomic context (AlphaGenome uses 1 Mb context because distal elements can influence genes from far away [73]). So another limitation is **contextual integration**: how do we place these synthetic sequences into a genome and ensure they work as intended? The DNA-Diffusion team addressed this partly by inserting sequences into known genomic loci and using a model to predict changes [69] [70]. They even found that inserting synthetic enhancers into previously inactive loci could *create* regulatory activity where there was none [11] [71], hinting that these sequences can impart function in new locations. However, the outcome might differ if multiple enhancers interact or if repressors in a cell also recognize the sequence. So, the limitation is that *generative models currently design sequences in a vacuum.* A future goal would be **co-**

designing sequences along with their genomic context or designing larger genomic constructs (like an entire gene with its regulatory domain). Evo 2's long context might assist here, but diffusion models might need to also scale up to longer ranges or work in tandem with predictive models that account for 3D genome context.

- **Limitations – Evaluation Metrics:** Both papers mention the lack of perfect metrics for sequence generation quality [56]. Unlike images where we have FID, or text where we can do human evaluation, for DNA we rely on proxies: k-mer statistics, motif enrichments, predictive model scores. These capture different aspects, but a model could potentially game one metric without truly being optimal (e.g., it could match motif frequency but scramble their order, resulting in non-functional sequence that still "looks" statistically fine). So there's a limitation in how we measure success. As generative DNA modeling is new, the community is still developing robust benchmarks. DiscDiff's introduction of S-FID (sequence Frechet Inception Distance using a neural embedding) is one attempt [26]. Over time, more nuanced evaluations, including **evolutionary conservation tests** (e.g., does a generated human sequence avoid something that would be deleterious, as judged by conservation or mutational constraint data?) or **physical property checks** (like DNA shape features), might be incorporated. Until then, the limitation is that a model might produce sequences that pass current tests but fail in aspects we didn't measure.

- **Applications – AI for Virtual Cells (AIVC):** The user mentions *AIVC (AI for Virtual Cells)*, which refers to the vision of having AI models that can simulate and design cellular behavior entirely in silico. Generative models like DNA-Diffusion and DiscDiff would be key components of such a system. Imagine a **"virtual cell"** where you have:

- A *predictive model* like **AlphaGenome** (DeepMind 2025) that can take a DNA sequence (up to 1 Mb) and predict everything about it – gene expression, chromatin state, 3D contacts, etc., across cell types [73].
- A *generative model* like DiscDiff that can propose new DNA sequence edits or entirely new constructs.
- One could loop these together in a design cycle: **use the generative model to propose a genomic alteration, then use the predictive model to simulate the cell's response**, then adjust the design accordingly. For instance, if you want to **"design a cell that produces insulin only when glucose is high"**, you could task the generative model to create a regulatory element that links a glucose-sensing pathway to the insulin gene, generate candidates, and then test them with the predictive model to see if indeed insulin expression goes up with glucose signals. This is speculative but illustrates how diffusion models could empower *in silico experimentation*. DNA-Diffusion's work of inserting sequences and checking gene activation with models [11] [71] is an early example of this paradigm.

- Another angle is **virtual screening of genomic variants**: generative models could sample plausible mutations or genomic variants an organism might have, and predictive models evaluate their impact on phenotype. This could help prioritize which mutations to test in disease research or which genomic edits might yield a desired trait in an engineered organism.

- **Creative and Evolutionary Insights:** Generative models might also help answer evolutionary questions: *"What might a gene's regulatory region look like if evolution had taken a different path?"* Since DiscDiff saw multi-species promoters, one could prompt it to generate a promoter for a human gene in the "style" of a zebrafish, for example, to see how sequence differences might still accomplish the

same function. This could yield insights into which parts of a sequence are functionally constrained and which can tolerate change. Similarly, generative models can be used to produce **negative examples** – sequences that are realistic but predicted to lack a certain function – which help to test the boundaries of what makes an enhancer vs. a non-enhancer.

In terms of **DiscDiff's "so-what"**: although the paper itself focuses on improving generative performance, the broader implication is that DiscDiff can be a backbone for many applications. By generating realistic DNA from multiple species, it could assist in **cross-species synthetic biology** (e.g. design a mouse sequence that mimics a human gene's regulation, aiding in creating better mouse models of human disease). It could also be extended for **controllable generation** – e.g., steering the latent diffusion to produce sequences with a specific property (they hinted Absorb-Escape allows some control, and one could imagine guiding diffusion by conditioning on metadata like "high expression" from the dataset) [74] [30]. Ultimately, DiscDiff and DNA-Diffusion showcase that **diffusion models unlock a lot of *potential creativity in genomics*** – we are no longer limited to sequences that evolution chanced upon; we can computationally explore countless "what-if" DNA designs.

In summary, generative diffusion models for DNA are powerful new tools with various possible applications: from designing gene therapies and synthetic organisms, to augmenting data for AI models, to testing biological hypotheses in silico. They are not without limitations – ensuring functional validity and integrating broader context remain challenges – but they mark a significant step toward *"Generative Genomics"*. By combining these models with advanced predictors (like Enformer or AlphaGenome) and eventually experimental feedback, we move closer to an era of **AI-assisted genome engineering**, where we can *ask* for a biological function and have algorithms propose DNA sequences to achieve it [3]. This synergy of generative and predictive models exemplifies the AIVC concept: a virtuous cycle where AI designs and evaluates virtual cell scenarios, accelerating our understanding and creation of biological systems in a safe, controlled manner. The work on DNA-Diffusion and DiscDiff not only provides two distinct perspectives on using diffusion in genomics, but also **inspires future research** to unify these perspectives – combining the fine control of conditional diffusion with the generality and scale of latent models. As techniques mature, we anticipate diffusion models to play a key role in the *"virtual cell" toolbox*, generating hypotheses and solutions that human imagination alone might not conceive.

Sources: The descriptions and comparisons above are based on the findings of Ferreira DaSilva *et al.* (2024) for DNA-Diffusion [1] [3] and Li *et al.* (2024) for DiscDiff [2] [21], as well as details from their methodology (e.g. the use of Gaussian noise in DNA-Diffusion [7] and the latent Autoencoder+diffusion in DiscDiff [39] [75]). DiscDiff's advantages in diversity and its hybrid Absorb-Escape approach are noted in the paper [49] [22]. The Evo 2 model is referenced from a 2025 preprint (Brixi *et al.*) which highlights its scale (9.3 Tb of data, 1M context) [43]. The evaluation and potential of these models are supported by the original authors' evaluations (e.g., DNA-Diffusion's use of Enformer-like predictors [11] and demonstration of induced gene accessibility [71], and DiscDiff's performance metrics on EPD-GenDNA [26]). Potential applications and future integration with predictive models like AlphaGenome [73] are extrapolated from the discussion in these works and the broader vision of AI in genomics. Overall, these sources collectively illustrate both the current accomplishments and the forward-looking possibilities for diffusion models in genomic sequence design.

---

1 3 4 10 DNA-DIFFUSION: LEVERAGING GENERATIVE MODELS FOR CONTROLLING CHROMATIN ACCESSIBILITY AND GENE EXPRESSION VIA SYNTHETIC REGULATORY ELEMENTS | OpenReview
https://openreview.net/forum?id=foByZC2svM

2 21 25 33 60 Paper page - DiscDiff: Latent Diffusion Model for DNA Sequence Generation
https://huggingface.co/papers/2402.06079

5 6 7 8 9 11 12 13 14 15 69 70 71 72 (PDF) DNA-Diffusion: Leveraging Generative Models for Controlling Chromatin Accessibility and Gene Expression via Synthetic Regulatory Elements
https://www.researchgate.net/publication/378007888_DNA-Diffusion_Leveraging_Generative_Models_for_Controlling_Chromatin_Accessibility_and_Gene_Expression_via_Synthetic_Regulatory_Elements

16 17 18 19 20 22 23 24 26 27 28 29 30 31 32 34 35 36 37 38 39 40 41 42 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 61 62 74 75 [2402.06079] DiscDiff: Latent Diffusion Model for DNA Sequence Generation
https://ar5iv.labs.arxiv.org/html/2402.06079

43 44 Genome Modeling Design Across All Domains of Life 2025.02.18.638918v1.full | PDF | Genetic Code | Gene
https://www.scribd.com/document/870754776/Genome-Modeling-Design-Across-All-Domains-of-Life-2025-02-18-638918v1-Full

63 64 65 66 67 68 arxiv.org
https://arxiv.org/pdf/2412.03278

73 AlphaGenome: advancing regulatory variant effect prediction with a unified DNA sequence model | Request PDF
https://www.researchgate.net/publication/393115584_AlphaGenome_advancing_regulatory_variant_effect_prediction_with_a_unified_DNA_sequence_model