A Survey of Discrete Diffusion for Text and Genomic Sequence Generation

Lars de Groot[1,2], Ruurd Kuiper[1], and Ayoub Bagheri[2]

[1] UMC Utrecht, Netherlands
L.A.deGroot-16@umcutrecht.nl, R.J.A.Kuiper@umcutrecht.nl
[2] Utrecht University, Netherlands
a.bagheri@uu.nl

Abstract. While diffusion models have achieved state-of-the-art results in continuous domains like image generation, their application to inherently discrete data such as natural language and DNA presents unique challenges. Continuous-space adaptations often introduce artifacts and complexities, motivating a focused investigation into models that operate directly on discrete data. This survey provides a comprehensive overview of the methods and advancements in the field of discrete diffusion models. We review the foundational formulations, including Denoising Diffusion Probabilistic Models (DDPMs) and Score-Based Generative Models (SGMs), and their theoretical adaptations to discrete state spaces. We then chronologically survey advancements across key modalities—Natural Language Processing and genomic sequences—examining critical research topics such as novel forward processes and the adaptation of pre-trained language models. By synthesizing these developments and outlining future research directions, this paper offers a structured overview to this rapidly evolving field.

1 Introduction

Generative modeling is a central task in machine learning that aims to learn a probability distribution from data, enabling the generation of novel samples. While various model families exist, including Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and flow-based models, diffusion models have emerged as a state-of-the-art paradigm, notably surpassing the previously dominant GANs in high-fidelity image generation Dhariwal and Nichol [2021]. Diffusion models were first introduced by Sohl-Dickstein et al. [2015], where theoretical formulations for both discrete and continuous data spaces were explored. However, subsequent research has been mainly focused on continuous data spaces with applications to image generation, leading to powerful models such as GLIDE, DALL-E 2, and Imagen, which have demonstrated remarkable success Conneau et al. [2020], Nichol et al. [2022], Ramesh et al. [2022]. Motivated by these results, research has increasingly focused on adapting these models to discrete data, which enables the modeling of sequences like natural language and genomic sequences Austin et al. [2021], Sarkar et al. [2024]. These methods can

be broadly grouped into two categories: *continuous diffusion*, which applies the diffusion process to continuous representations (e.g., embeddings) Gulrajani and Hashimoto [2023] that stand in for discrete values, and *discrete diffusion* Austin et al. [2021], which operates directly on the discrete values themselves. While both approaches have been explored, initial research often favored continuous diffusion formulations due to the availability of established methods from image generation. However, in fields like natural language processing (NLP) and genomic sequence modeling. It has been found that continuous formulations can suffer from higher sampling time and may exhibit "rounding errors" when mapping output embeddings back to their corresponding discrete tokens Lou et al. [2024], Li et al. [2024]. Consequently, subsequent research has increasingly shifted focus toward discrete formulations. This rapid proliferation of distinct methods, theoretical insights, and novel applications within the discrete diffusion landscape has created a need for a structured review to integrate these developments.

In this survey, we explore the advancements and research directions of discrete diffusion models. To best address inherently discrete data types like natural language and genomic sequences, we focus exclusively on fully discrete diffusion models, setting aside their continuous counterparts, while referring the reader to existing surveys on continuous diffusion models Li et al. [2023], Yi et al. [2024], Zou et al. [2023], Zhu and Zhao [2023]. We will examine the main theoretical foundations of diffusion models, key research interests, and summarize advancements within each modality. We will also provide a chronological perspective on the development of these models across different modalities, reflecting on the findings and highlighting potential future research directions. Our research questions are as follows:

- What advancements have been made for discrete diffusion models?
- How have formulations for continuous diffusion models been adapted to the discrete data domain?
- What research directions have been explored within the field?
- What are the potential research directions for natural language and DNA generating using diffusion models?

The rest of the paper is structured as follows. Section 2 introduces a framework to understand diffusion models. Section 3 details the different noising processes. Section 4 discusses methods to adapt autoregressive models to use diffusion. Section 5 outlines a chronological evolution of research results and directions per modality and 6 discusses limitations and future research directions.

2 Formulations of diffusion models

Diffusion models are best understood within the broader context of generative modeling. The goal of generative modeling is to create a model that can generate novel samples that are characteristic of the data distribution on which it was

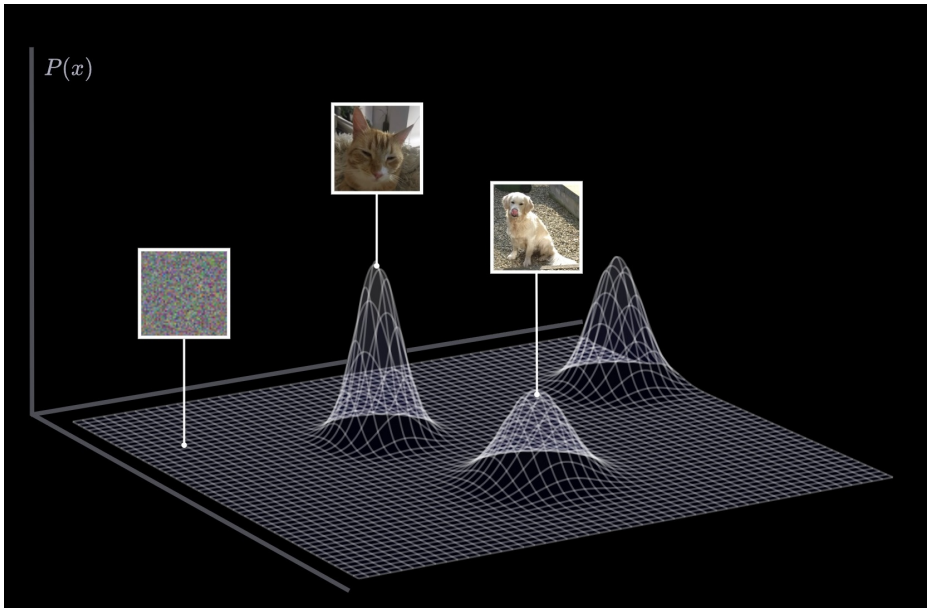

Fig. 1. A simplified visualization of the probability distribution of image data. The surface represents the probability density landscape, where the peaks correspond to regions of high probability containing meaningful data, such as the images of the cat and dog. The flat 'valleys' represent regions of low probability, where data points are noise.

trained. In this sense, the model learns an approximation of the probability distribution of the training data Song et al. [2021]. This concept can be visualized as a landscape shaped by probability; hills represent regions of high probability where data points are concentrated, while valleys represent low-probability regions, which can be seen as pure noise as shown in Figure 1. Within this framework, a diffusion model learns to navigate from these valleys to the peaks, which represent plausible data points. It achieves this by taking small, iterative steps across the data space, with each step designed to gradually increase the probability of its position.

An analogy is that the model learns to act as a compass. At any given point in the landscape, it indicates the direction of steepest ascent in probability, guiding the generation process step-by-step from noise towards a high-likelihood data sample. To learn this navigation, the model is trained on data at various stages of noising. This is achieved through the forward process, where a predefined schedule of noise is incrementally added to the training data. This process gradually transforms a data point (x_0) into pure, unstructured noise (x_T), as shown in Figure 2. The model's task is then to learn the reverse process: given a noisy data point (x_T) it is trained to remove noise over multiple small iterative

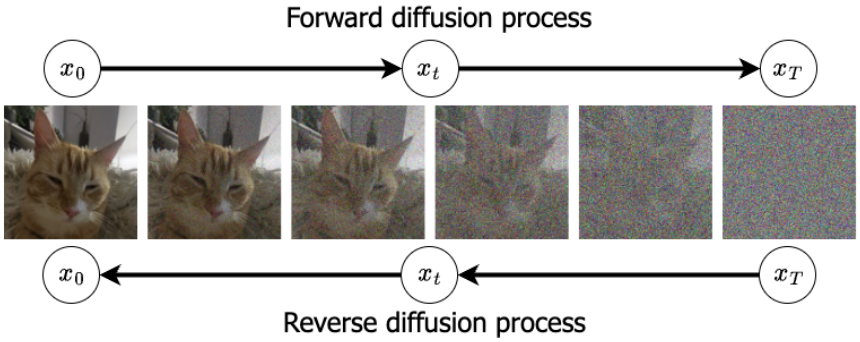

Fig. 2. An intuitive example of the diffusion process. The forward process gradually corrupts data (x_0) by adding noise over several timesteps (x_t), eventually resulting in pure, unstructured noise (x_T). A model then learns the reverse process, which starts with noise and learns to denoise it back into a clean sample.

time steps, until it reaches a plausible data point (x_0). Thus, the model learns to reverse the diffusion process.

Although all diffusion models are based on this forward-reverse process paradigm, they differ in their mathematical formulations. These formulations are not mutually exclusive; rather, they represent different methods of defining the same core diffusion process. Consequently, there is not one that always outperforms the other, but their different mathematical formulation do enable different sampling techniques and training objectives which come with their own characteristics.

So far, we have intuitively described this diffusion as occurring over a series of discrete timesteps, as labeled from x_0 to x_T in Figure 2. However it is also possible to formulate diffusion models with a continuous time variable. Empirical results indicate that continuous-time approaches can offer slight performance advantages over their discrete-time counterparts, particularly in terms of perplexity and generation quality, though these improvements are generally modest Sahoo et al. [2024].

There are many established formulations of diffusion models, the most known of them are: denoising diffusion probabilistic models (DDPMs) Ho et al. [2020], Score-based Generative Models (SGMs) Song et al. [2021], Denoising Diffusion Implicit Models (DDIM) Song et al. [2022], diffusion models using Stochastic Differential Equations (SDEs) or Odinary Differential Equations (ODE) Song et al. [2021].

While all these formulations were originally made to model continuous data, DDPM and SGMs have since been adapted to work directly on discrete data Austin et al. [2021], Meng et al. [2023].We give an introduction to these formulations and how they have been adapted to work on discrete data in appendix 8.2.

3 Forward processes

Diffusion models consist of a forward process that incrementally perturbs data and a learned reverse process designed to denoise the data by undoing these perturbations Johnson et al. [2021]. As a result, changes to the forward process redefine the task the model learns and can yield desirable reverse processes.

The foundational work of Forre et al. [2022] uses a forward process with uniform resampling. In this approach, the probability that each token in a sequence is uniformly resampled increases as the time step t approaches T. At $t = T$, the sequence x_T ideally follows a uniform distribution over all possible tokens.

Austin et al. [2021] introduced a framework for discrete diffusion models, utilizing transition matrices to define various forward processes. A notable example is *absorbing diffusion*, where the probability of each token transitioning to a special [MASK] token increases with t. In this scheme, the fully corrupted sequence x_T consists entirely of [MASK] tokens. This approach shares conceptual similarities with Masked Language Modeling (MLM) Devlin et al. [2019].

In the same work Austin et al. [2021] also proposed a forward process based on token embedding distance. This method corrupts tokens by swapping them with others, with a higher probability of swapping to tokens that are semantically or syntactically similar (i.e., closer in the embedding space). Despite this structured noise, the process is designed such that x_T still converges to a uniform distribution of tokens.

The aforementioned forward processes typically modify sequence elements inplace. This constraint can pose challenges for the denoising model; for instance, if an element is predicted in an incorrect position, subsequent denoising steps may become more difficult. To address this, Johnson et al. [2021] incorporated explicit insertion [INS] and deletion [DEL] operations into the corruption and denoising framework. This allows the model to insert new elements between existing ones or remove elements entirely, providing greater flexibility to correct erroneously placed tokens during generation.

He et al. [2022] developed the spindle noise schedule, which can be viewed as an adaptive forward process influencing not only the rate but also the order of token corruption. This schedule addresses two key observations: (1) different tokens within a sequence convey varying amounts of information, and (2) denoising language models often predict high-frequency (less surprising) tokens earlier in the reverse process, thereby rapidly increasing likelihood. The spindle noise schedule corrupts sequences by ensuring a uniform amount of information is degraded at each step and by corrupting the most informative tokens early in the forward process. This design encourages the model to generate less informative ("easy") tokens first during its backward (denoising) process He et al. [2022].

4 Diffusion adaptation

The similarity between the training objective of MLMs and the denoising task in masked diffusion models was notably explored in Austin et al. [2021]. MLMs,

such as BERT, are pretrained using a bidirectional attention mechanism, allowing each token to attend to all other tokens in the sequence to predict randomly masked positions. This inherent bidirectionality made them suitable candidates for adaptation into denoising networks within diffusion frameworks, a strategy leveraged by subsequent works such as He et al. [2022] and Ye et al. [2023]. Given the wide availability of pretrained language models (pLMs) and the significant cost of pre-training new models from scratch, adapting existing pLMs for diffusion-based generation is a compelling research direction.

A key difference in their pretraining objectives lies in the scope of the denoising task: MLMs are typically trained by masking a fixed, small percentage of tokens (e.g., 15%), whereas a diffusion process involves a spectrum of masking levels, from 0% to 100% masked tokens. Consequently, these MLM-based models generally require fine-tuning to effectively serve as denoising networks across the full range of diffusion timesteps.

However, many of the largest and most capable language models are autoregressive (AR) causal language models, employing causal attention masks. These masks restrict the attention mechanism, permitting a token to attend only to preceding tokens in the sequence. This fundamental architectural difference poses a challenge when adapting AR models to diffusion frameworks, which often benefit from or require bidirectional context for optimal denoising. The work by Gong et al. [2024] addresses this disparity. They introduce "attention mask annealing"—a technique to gradually transition the model from its native causal attention to a more bidirectional attention suitable for diffusion—alongside a mechanism to handle the "shift operation" (aligning the AR model's next-token prediction with the diffusion model's current-token denoising objective), thereby enabling the adaptation of causal language models to a diffusion-based generative framework.

A primary motivation for repurposing pLMs is to leverage the representations learned during their pre-training. This approach has proven effective, with Gong et al. [2024] showing that diffusion models adapted from AR foundations can achieve performance competitive with their original AR counterparts across various tasks. While these adapted models do not represent a universal improvement—exhibiting stronger performance on some benchmarks and weaker on others—the value of knowledge transfer is evident. A direct comparison by Han et al. [2024] revealed that adapted diffusion models outperform their randomly initialized counterparts. This finding strongly suggests that the adaptation process successfully repurposes the learned representations of the original autoregressive models, providing a more effective starting point than training from scratch.

Notably adapting AR pretrained models to diffusion has been explored for a very limited extent in the field of genomics. Penzar et al. [2023] adapts a sequence to expression model to diffusion and uses it to generate DNA sequences.

5 Advancements across modalities

This section examines the motivations, key results, and major research topics for discrete diffusion models across two principal modalities: Natural language and DNA. We provide a chronological overview of the field's progression to illustrate how it has reached its current state. Subsequently, we elaborate on individual models, summarizing their core methodologies and significant findings. Table 1 in Appendix 8.1 summarizes all diffusion models for NLP covered in this survey, while Table 2 similarly summarizes the diffusion models for genomic sequence modeling.

5.1 Natural language

Generative modeling has become widespread in the field of NLP. With applications of these models now in daily public use, advancements in this domain are of significant interest. Consequently, the potential for diffusion models to serve as an alternative to the dominant autoregressive paradigm has attracted considerable research attention Ye et al. [2023]Deschenaux and Gulcehre [2024]Sahoo et al. [2024]He et al. [2022] Lou et al. [2024]Ye et al. [2024b]Gulrajani and Hashimoto [2023]Wu et al. [2023]Gong et al. [2023]Zhang et al. [2024].

Diffusion models possess several distinct advantages over autoregressive models. A primary advantage is their capacity for parallel token generation, which can lead to increased sampling speed and efficiency. Furthermore, their potential for sequence-agnostic processing allows for a bidirectional context window, addressing a known limitation of autoregressive (AR) causal language models. This unidirectional constraint in AR models is implicated in phenomena such as the "reversal curse," where a model learns a directional relationship (e.g., "A is B") but fails to infer its converse ("B is A") Berglund et al. [2024].

Initial research into natural language generation with diffusion models often involved applying the diffusion processes to continuous token embedding spaces Gulrajani and Hashimoto [2023]Wu et al. [2023]Gong et al. [2023]Zhang et al. [2024]. This approach allowed for the adaptation of continuous diffusion methods, typically developed for image generation. However, continuous-space diffusion for natural language can necessitate a higher number of diffusion steps (and network evaluations), as multiple small perturbations in the embedding space may be required to effect a change equivalent to altering a single discrete token Lou et al. [2024]. Additionally, this continuous approach requires a dequantization step to map the diffused token embeddings back to discrete tokens, a non-trivial process that significantly impacts model performance Li et al. [2022]Lin et al. [2022]Shabalin et al. [2025].

The viability of fully discrete diffusion models as generative language models prompted investigations into their scalability Nie et al. [2024]. Given that the success of Large Language Models (LLMs) is significantly attributed to their massive parameter counts Brown et al. [2020] Hoffmann et al. [2022], comparable scaling properties are crucial for discrete diffusion models to be competitive

alternatives. Nie et al. [2024] found that masked diffusion models (MDMs) exhibit scaling laws similar to AR models, although they required approximately 16 times more compute to achieve the same validation loss. This compute gap, however, might be narrowed through optimizations in training and sampling procedures. Subsequent work by Nie et al. [2025] continued to explore the scaling of diffusion language models (DLMs). They contended that validation loss alone might not fully represent model performance, instead evaluating models on a suite of language understanding and reasoning benchmarks such as MMLU, ARC-C, and PIQA. Their findings indicated that DLMs matched the scaling trends of AR models on these tasks, and in some instances, even surpassed them Nie et al. [2025].

Nie et al. [2024] also investigated whether DLMs could overcome the "reversal curse" Berglund et al. [2024], a phenomenon previously discussed where AR models struggle with bidirectional inference due to their unidirectional attention mechanism. In contrast, diffusion models typically employ a bidirectional attention mechanism, allowing tokens to attend to the entire context. Nie et al. [2024] reported that their diffusion models were largely able to overcome the reversal curse, achieving high accuracy on inferring reversed relationships where AR models failed. The ability of diffusion models to overcome the reversal curse highlights a potential fundamental difference in how they process and reason about information compared to AR models. Ye et al. [2023] explored this further by adapting MLMs like XLM-RoBERTa for diffusion-based generation and evaluating their reasoning capabilities on the GSM8K benchmark, which comprises grade-school math problems requiring multi-step reasoning. While their models did not match the performance of strong AR models on these complex reasoning tasks, qualitative analysis revealed that the diffusion models exhibited distinct reasoning patterns, including forms of backward reasoning and an inclination to generate final answers early in the process. Building on the concept of explicit reasoning steps, Ye et al. [2024b] introduced Diffusion-of-Thought (DoT), an analogue to the Chain-of-Thought (CoT) prompting used in AR models. Unlike CoT, which generates reasoning tokens sequentially, DoT integrates these thinking steps within the reverse diffusion process. Evaluations on GSM8K showed that DoT-enabled diffusion models could outperform CoT-prompted AR models of similar size. Furthermore, DoT demonstrated a practical trade-off between efficiency and performance, with accuracy improving as the number of diffusion steps increased.

5.2 Genomic sequences

The challenge of modeling genomic sequences shares fundamental properties with NLP, as both revolve around generating sequences of discrete elements. This parallel has historically justified the transfer of methods from one field to the other, with autoregressive models having been applied extensively to the genomic sequence modeling task Benegas et al. [2025]. Following this precedent, diffusion models now represent a promising new approach for this cross-domain application. However, modeling genomic sequences presents unique challenges

distinct from natural language, primarily stemming from its sparsity and the intricate nature of genomic regulation Sarkar et al. [2024]. A significant portion of the genome is non-coding, and within this, only a small fraction comprises cis-regulatory elements (e.g., transcription factor binding sites, promoters) crucial for gene expression. This creates a substantial class imbalance between functionally informative and non-informative positions, making the accurate modeling of DNA's regulatory grammar an ongoing research frontier Sarkar et al. [2024].

While autoregressive genomic language models (GLMs) have been explored Nguyen et al. [2023] Benegas et al. [2025], the bidirectional contextual understanding offered by diffusion models presents a compelling alternative, particularly given their success in other domains Yang et al. [2023]. Current research in applying diffusion models to DNA predominantly focuses on the conditional generation of regulatory sequences. This often involves leveraging guidance mechanisms to generate DNA with desired characteristics, such as specific expression levels, cell-type specificity, or species-specific features.

An early approach in this direction was the Dirichlet Diffusion Score Model (DDSM) Avdeyev et al. [2023]. DDSM introduced a score-based Stochastic Differential Equation (SDE) diffusion model operating in continuous probability simplex space. A key innovation was its method for handling discrete DNA data: the diffusion process was defined such that its stationary distribution is the Dirichlet distribution, providing a natural way to map the model's continuous outputs back to discrete nucleotide probabilities. Conditioned on transcription initiation signal profiles—which quantify the frequency of transcription initiation at each nucleotide position—DDSM was capable of generating human promoter sequences designed to exhibit specific transcription initiation patterns, thereby offering a means to control predicted expression levels.

Subsequently, DNA-Diffusion DaSilva et al. [2024] employed a Denoising Diffusion Probabilistic Model (DDPM) formulation, also operating on continuous values. Unlike DDSM's simplex space, DNA-Diffusion represented DNA sequences as one-hot encoded vectors with binary values mapped to a continuous range of [-1, 1]. Diffusion and denoising occurred in this continuous space, with the final discrete sequence recovered by taking the argmax over the channel dimension at each position. Conditioned on cell type, DNA-Diffusion generated sequences that were distinct from the training set and, importantly, exhibited predicted increases in regulatory activity within the target cell type.

A shift towards models operating directly on discrete data was marked by DNA Discrete Diffusion (D3) Sarkar et al. [2024]. Adopting the formulation for discrete diffusion based on estimating data distribution ratios (as in Lou et al. [2024], foundational to D3's approach), D3 represented a significant step as a fully discrete DNA diffusion model. Its forward process was specifically tailored for genomic sequences, designed to converge to a uniform distribution and perturbing only a single nucleotide at each step. Evaluated on diverse tasks including human promoter, fly enhancer, and cell-type-specific regulatory sequence generation, D3 demonstrated an improved ability to capture the diversity of cis-regulatory grammars compared to previous methods Sarkar et al. [2024].

Another distinct strategy, Latent Diffusion Models (LDMs), was introduced to DNA sequence generation by DiscDiff Li et al. [2024]. This approach first encodes DNA sequences into a continuous latent space using a Variational Autoencoder (VAE). Diffusion and denoising processes are subsequently performed within this lower-dimensional latent space, after which the VAE's decoder reconstructs novel DNA sequences from the generated latent representations. Recognizing that LDM-generated sequences may exhibit errors at the single nucleotide level, DiscDiff also proposes 'Absorb-Escape.' This post-processing technique utilizes an autoregressive model to resample specific nucleotides that were assigned low confidence by the diffusion model during generation.

The paradigm of masked discrete diffusion was applied to DNA sequence modeling by Sahoo et al. in their work on Masked Diffusion Language Models (MDLM) Sahoo et al. [2024], primarily focused on natural language but with notable experiments on DNA. MDLM is formulated as a discrete diffusion model employing an absorbing state (masking) forward process and a continuous-time Rao-Blackwellized ELBO that simplifies to a weighted average of masked language modeling losses. This work marked a key application of absorbing state diffusion to DNA. Notably, their MDLM, when applied to DNA, achieved generative perplexity approaching that of HyenaDNA Nguyen et al. [2023], a prominent autoregressive DNA language model. This result underscored the potential of discrete diffusion models, particularly those with efficient objectives like MDLM, as competitive alternatives for DNA sequence modeling.

6 Discussion and Future Directions

6.1 DNA diffusion models

The application of diffusion models to DNA sequence modeling, while promising, remains less explored compared to their use in NLP. Yet, diffusion models possess characteristics that could be uniquely advantageous for DNA generation. The bidirectional context inherent in DNA, and the capacity of diffusion models to capture complex distributions and generate diverse, novel outputs, make them a compelling avenue for generating functional DNA sequences.

A key research direction that garners attention is *increasing the sequence length of these models*. Longer sequence contexts are crucial for accurately modeling interactions between distal cis-regulatory elements, a known challenge in genomics Sarkar et al. [2024]. Successfully scaling sequence length could enable applications such as the de novo design of DNA inserts that are contextually aware of their surrounding genomic environment. Diffusion models, with their inherent ability for infilling and conditional generation, are particularly well-suited for such tasks.

Multitask learning with diffusion models in genomics is another promising, yet relatively unexplored, frontier. This approach involves training a single model to concurrently predict different types of genomic data, such as DNA sequence, associated gene expression values, and chromatin accessibility. While the intuition is that learning shared representations across related tasks could improve model performance or data efficiency Kathail et al. [2024], comprehensive evidence demonstrating consistent, significant gains in predictive performance over single-task models across diverse genomic applications is still developing. Nevertheless, it stands to reason that leveraging the interdependencies between different genomic modalities could lead to more robust predictions. Such multi-task models could also enable precise control over sequence generation, for instance, by specifying a target expression profile and generating the corresponding DNA sequence. A significant hurdle, however, is the limited availability of large-scale, matched multi-modal genomic datasets, which currently constrains the training of these models.

6.2 Diffusion adaptation

The similarities between the MLM objective and the denoising process in diffusion models have enabled the adaptation of pretrained language models for diffusion tasks. However, *open questions remain regarding the comparative performance of these 'adapted diffusion models' versus 'true diffusion models' trained from scratch*. Notably, for natural language modelling, it has not been tested if adapted diffusion models beat the reversal curse to the same extent that true diffusion models do.

This leads to a crucial question: *do adapted diffusion models acquire new knowledge during the adaptation process, or do they primarily repurpose existing*

representations? Investigating the extent of parameter changes during adaptation, similar to analyses done for reinforcement learning Mukherjee et al. [2025], could offer insights into how drastically model weights are altered. However, a more qualitative experiment would be needed to fully asses if the adaptation to diffusion results in new knowledge, like testing if previous unknown relationships have been inferred by the adaptation to diffusion.

Furthermore, while studies He et al. [2022], Gong et al. [2024] describe general training regimes for adaptation (e.g., number of tokens and training duration), the reasoning determining the optimal amount of data or training duration for effective adaptation often come down to, convergence on validation metrics or efficient use of computational resources. While effective, such approaches may not be available in compute-constrained scenarios. This motivates the need for more efficient training schemes, such as those leveraging parameter-efficient fine-tuning methods like LoRA Hu et al. [2021]. A first example of this approach is LoRA-Adapted Diffusion (LAD), which demonstrates that LoRA fine-tuning alone can be sufficient to transform a pretrained autoregressive model into an effective diffusion-based generator Kuiper et al. [2025]. However the use of those methods has only been explored to a limited extent Gong et al. [2024], and questions remain about it's impact on the adaptation to diffusion.

At the time of writing, diffusion adaptation has not been applied in the field of DNA modelling. A key consideration for this adaptation is the diversity of tokenization strategies used by AR DNA language models. While the DNA diffusion models explored in this survey operate at the base-pair level, existing AR models use strategies based on either k-mers or individual base pairs Benegas et al. [2025]. However, there is no fundamental reason why diffusion models could not also use k-mer tokenization. Furthermore, if we generalize from findings in diffusion for natural language generation, fully discrete techniques may be a more natural fit for k-mer vocabularies than the continuous methods explored so far. Nonetheless, adapting AR DNA language models for diffusion remains an unexplored avenue of research.

7 Conclusion

Diffusion models are establishing themselves as competitive alternatives to autoregressive methods for generating discrete sequential data. Having demonstrated their potential beyond their origins in image generation, their successful adaptation to the discrete data space has unlocked new possibilities for modeling complex data distributions like natural language and genomic sequences. This survey has charted the field's progression from continuous-space adaptations toward fully discrete formulations, which have resolved key challenges such as embedding rounding errors and inefficient sampling. Our review highlights several critical advancements. The development of diverse forward processes, particularly those based on absorbing/masking states, has created a strong conceptual link to MLMs. This link has enabled a significant research direction: the adaptation of pretrained language models for diffusion-based generation. In

natural language generation, diffusion models have shown interesting characteristics, such as beating the "reversal curse," hinting at improved reasoning. In genomics, while still a nascent field of application, diffusion models are demonstrating a critical application in the conditional generation of functional DNA sequences, moving beyond simple generation to guided molecular design.

Despite these successes, much work lies ahead. While diffusion models are proving their competitiveness, they do not yet represent a universal improvement over existing methods, with challenges like the increased training compute required for scaling remaining an open area for optimization. We are optimistic that discrete diffusion models can achieve a foundational role in both natural language and genomics, but further efforts are needed to rigorously benchmark their capabilities and demonstrate their utility. As we have discussed, future research should not only focus on optimizing these methods but also on tackling fundamental questions about the nature of the knowledge they acquire and the unique tasks for which they are best suited.

Acknowledgments

This research was supported by the Utrecht University focus area Applied Data Science. We thank the ADS steering committee for their support.

Bibliography

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured Denoising Diffusion Models in Discrete State-Spaces. *ArXiv*, July 2021.

Pavel Avdeyev, Chenlai Shi, Yuhao Tan, Kseniia Dudnyk, and Jian Zhou. Dirichlet Diffusion Score Model for Biological Sequence Generation, June 2023.

Gonzalo Benegas, Chengzhong Ye, Carlos Albors, Jianan Canal Li, and Yun S. Song. Genomic language models: Opportunities and challenges. *Trends in Genetics*, 0(0), January 2025. ISSN 0168-9525. https://doi.org/10.1016/j.tig.2024.11.013.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A", May 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, July 2020.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A Continuous Time Framework for Discrete Denoising Models, October 2022.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale, April 2020.

Lucas Ferreira DaSilva, Simon Senan, Zain Munir Patel, Aniketh Janardhan Reddy, Sameer Gabbita, Zach Nussbaum, César Miguel Valdez Córdova, Aaron Wenteler, Noah Weber, Tin M. Tunjic, Talha Ahmad Khan, Zelun Li, Cameron Smith, Matei Bejan, Lithin Karmel Louis, Paola Cornejo, Will Connell, Emily S. Wong, Wouter Meuleman, and Luca Pinello. DNA-Diffusion: Leveraging Generative Models for Controlling Chromatin Accessibility and Gene Expression via Synthetic Regulatory Elements, February 2024.

Justin Deschenaux and Caglar Gulcehre. Beyond Autoregression: Fast LLMs via Self-Distillation Through Time, October 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019.

Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis, June 2021.

P. Forre, E. Hoogeboom, P. Jaini, D. Nielsen, and M. Welling. *Argmax Flows and Multinomial Diffusion: Learning Categorical Distributions*. 15San Diego, CANeural Information Processing Systems Foundation, 2022. ISBN 978-1-7138-4539-3.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. DiffuSeq: Sequence to Sequence Text Generation with Diffusion Models, February 2023.

Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, Hao Peng, and Lingpeng Kong. Scaling Diffusion Language Models via Adaptation from Autoregressive Models, October 2024.

Ishaan Gulrajani and Tatsunori B. Hashimoto. Likelihood-Based Diffusion Language Models, May 2023.

Kehang Han, Kathleen Kenealy, Aditya Barua, Noah Fiedel, and Noah Constant. Transfer Learning for Text Diffusion Models, January 2024.

Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. DiffusionBERT: Improving Generative Masked Language Models with Diffusion Models, November 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. https://arxiv.org/abs/2006.11239v2, June 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models, March 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021.

Aapo Hyvarinen. Estimation of Non-Normalized Statistical Models by Score Matching. page 15, 2005.

Aapo Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51(5):2499–2512, February 2007. ISSN 01679473. https://doi.org/10.1016/j.csda.2006.09.003.

Aapo Hyvarinen. Connections Between Score Matching, Contrastive Divergence, and Pseudolikelihood for Continuous-Valued Variables. *IEEE Transactions on Neural Networks*, 18(5):1529–1531, September 2007. ISSN 1045-9227. https://doi.org/10.1109/TNN.2007.895819.

Daniel D. Johnson, Jacob Austin, Rianne van den Berg, and Daniel Tarlow. Beyond In-Place Corruption: Insertion and Deletion In Denoising Probabilistic Models, July 2021.

Pooja Kathail, Ayesha Bajwa, and Nilah M. Ioannidis. Leveraging genomic deep learning models for non-coding variant effect prediction, November 2024.

Hyukhun Koh, Minha Jhang, Dohyung Kim, Sangmook Lee, and Kyomin Jung. PLM-Based Discrete Diffusion Language Models with Entropy-Adaptive Gibbs Sampling, December 2024.

Ruurd Jan Anthonius Kuiper, Maarten van Smeden, Lars de Groot, Bram van Es, and Ayoub Bagheri. LAD: LoRA-Adapted Diffusion. In *The 2025 Conference on Empirical Methods in Natural Language Processing - System Demonstrations*, September 2025.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-LM Improves Controllable Text Generation, May 2022.

Yifan Li, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Diffusion Models for Non-autoregressive Text Generation: A Survey, May 2023.

Zehui Li, Yuhao Ni, William A V Beardall, Guoxuan Xia, Akashaditya Das, Guy-Bart Stan, and Yiren Zhao. DiscDiff: Latent Diffusion Model for DNA Sequence Generation. 2024. https://doi.org/10.48550/ARXIV.2402.06079.

Zheng-Wen Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Weizhu Chen, and Nan Duan. GENIE : Large Scale Pre-training for Generation with Diffusion Model. 2022.

Sulin Liu, Juno Nam, Andrew Campbell, Hannes Stärk, Yilun Xu, Tommi Jaakkola, and Rafael Gómez-Bombarelli. Think While You Generate: Discrete Diffusion with Planned Denoising, October 2024.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution, June 2024.

Siwei Lyu. Interpretation and Generalization of Score Matching, May 2012.

Chenlin Meng, Kristy Choi, Jiaming Song, and Stefano Ermon. Concrete Score Matching: Generalized Score Matching for Discrete Data, January 2023.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tur, and Hao Peng. Reinforcement Learning Finetunes Small Subnetworks in Large Language Models, May 2025.

Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Callum Birch-Sykes, Michael Wornow, Aman Patel, Clayton Rabideau, Stefano Massaroli, Yoshua Bengio, Stefano Ermon, Stephen A. Baccus, and Chris Ré. HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution, November 2023.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, March 2022.

Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. Scaling up Masked Diffusion Models on Text, October 2024.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large Language Diffusion Models, February 2025.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your Absorbing Discrete Diffusion Secretly Models the Conditional Distributions of Clean Data, July 2024.

Dmitry Penzar, Daria Nogina, Elizaveta Noskova, Arsenii Zinkevich, Georgy Meshcheryakov, Andrey Lando, Abdul Muntakim Rafi, Carl de Boer, and Ivan V Kulakovskiy. LegNet: A best-in-class deep learning model for short

DNA regulatory regions. *Bioinformatics*, 39(8):btad457, August 2023. ISSN 1367-4811. https://doi.org/10.1093/bioinformatics/btad457.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents, April 2022.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and Effective Masked Diffusion Language Models, November 2024.

Anirban Sarkar, Ziqi Tang, Chris Zhao, and Peter K. Koo. Designing DNA With Tunable Regulatory Activity Using Discrete Diffusion, May 2024.

Alexander Shabalin, Viacheslav Meshchaninov, Egor Chimbulatov, Vladislav Lapikov, Roman Kim, Grigory Bartosh, Dmitry Molchanov, Sergey Markov, and Dmitry Vetrov. TEncDM: Understanding the Properties of the Diffusion Model in the Space of Language Model Encodings, February 2025.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and Generalized Masked Diffusion for Discrete Data, January 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265. PMLR, June 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models, October 2022.

Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution, October 2020.

Yang Song, Jascha Narain Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. *ArXiv*, November 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations, February 2021.

Pascal Vincent. A Connection Between Score Matching and Denoising Autoencoders. *Neural Computation*, 23(7):1661–1674, July 2011. ISSN 0899-7667, 1530-888X. https://doi.org/10.1162/NECO_a_00142.

Tong Wu, Zhihao Fan, Xiao Liu, Yeyun Gong, Yelong Shen, Jian Jiao, Hai-Tao Zheng, Juntao Li, Zhongyu Wei, Jian Guo, Nan Duan, and Weizhu Chen. AR-Diffusion: Auto-Regressive Diffusion Model for Text Generation, December 2023.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Comput. Surv.*, 56 (4):105:1–105:39, November 2023. ISSN 0360-0300. https://doi.org/10.1145/3626235.

Jiacheng Ye, Jiahui Gao, Shansan Gong, Lin Zheng, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Beyond Autoregression: Discrete Diffusion for Complex Reasoning and Planning, October 2024a.

Jiacheng Ye, Shansan Gong, Liheng Chen, Lin Zheng, Jiahui Gao, Han Shi, Chuan Wu, Xin Jiang, Zhenguo Li, Wei Bi, and Lingpeng Kong. Diffusion of Thoughts: Chain-of-Thought Reasoning in Diffusion Language Models, December 2024b.

Jiasheng Ye, Zaixiang Zheng, Yu Bao, Lihua Qian, and Quanquan Gu. Diffusion Language Models Can Perform Many Tasks with Scaling and Instruction-Finetuning, August 2023.

Qiuhua Yi, Xiangfan Chen, Chenwei Zhang, Zehai Zhou, Linan Zhu, and Xiangjie Kong. Diffusion models in text generation: A survey. *PeerJ Computer Science*, 10:e1905, February 2024. ISSN 2376-5992. https://doi.org/10.7717/peerj-cs.1905.

Yizhe Zhang, Jiatao Gu, Zhuofeng Wu, Shuangfei Zhai, Josh Susskind, and Navdeep Jaitly. PLANNER: Generating Diversified Paragraph via Latent Language Diffusion Model, March 2024.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked Diffusion Models are Secretly Time-Agnostic Masked Models and Exploit Inaccurate Categorical Sampling, February 2025.

Yuansong Zhu and Yu Zhao. Diffusion Models in NLP: A Survey, March 2023.

Hao-Li Zou, Zae Myung Kim, and Dongyeop Kang. A Survey of Diffusion Models in Natural Language Processing. May 2023.

8 Appendix

8.1 Tables

Table 1. Summary of discrete diffusion models for natural language generation. The column *Type* indicates if a model uses a discrete (D) or continuous (C) formulation and the column *Time* indicates if a model uses a discrete time variable (D) or a continuous time variable (C).

Model Name	Type	Formulation	Time	Forward Process	Sampling Technique	Noising Schedule
DiffusionBERT He et al. [2022]	D	DDPM	D	Absorbing	Not specified	Spindle
D3PM Austin et al. [2021]	D	DDPM	D	Uniform, Absorbing, Discretized Gaussian, Token embedding distance	Not specified	Linear, Cosine
XLM-R[a]Ye et al. [2023]	D	DDPM	D	Absorbing	Not specified	Spindle
SEDD Lou et al. [2024]	D	SGM	C	Uniform, Absorbing	Euler method, Tweedie τ-leaping	Geometric, log-linear
RADD Ou et al. [2024]	D	SGM	C	Absorbing	Euler method, Tweedie τ-leaping	Log-linear
DDPD Liu et al. [2024]	D	DDPM	C	Uniform, Absorbing	Adaptive Gillespie	Not specified
MDM Ye et al. [2024a]	D	DDPM	D	Absorbing	easy-first TopK decoding	Not specified
DiffuGPT, DiffuLLaMA Gong et al. [2024]	D	DDPM	C	Absorbing	Ancestral sampling	Linear
MDM Nie et al. [2024]	D	DDPM	C	Absorbing	Greedy sampling	Not specified
MDLM Sahoo et al. [2024]	D	DDPM	C	Absorbing	Ancestral sampling, Semi-Autoregressive	Log-linear
Diffusion-EAGS Koh et al. [2024]	D	DDPM	D	Entropy-based noising	Entropy-Adaptive Gibbs Sampling	Entropy-based
MD4, GenMD4 Shi et al. [2025]	D	DDPM	C	Absorbing	Ancestral sampling	Linear, Geometric, Cosine
MDM Zheng et al. [2025]	D	DDPM	C	Absorbing	First-Hitting Sampler	Not specified
LLaDA Nie et al. [2025]	D	DDPM	C	Absorbing	Low-confidence remasking, semi-AR remasking	Warmup-Stable-Decay

[a] The paper adapted an existing XLM-RoBERTa model as the denoising network.

Table 2. Summary of diffusion models for DNA sequence generation. The column *Type* indicates the diffusion space: Discrete (D) or Continuous (C). The column *Time* indicates if a model uses a discrete time variable (D) or a continuous time variable (C).

Model Name	**Type**	**Formulation**	**Time**	**Forward Process**	**Sampling Technique**	**Noising Schedule**
DDSM Avdeyev et al. [2023]	C	SGM / SDE	C	Multivariate Jacobi	Time-dilation	Not applicable (uses weighting function)
DNA-Diffusion DaSilva et al. [2024]	C	DDPM	D	Gaussian noise	Iterative denoising	Linear
D3 Sarkar et al. [2024]	D	SGM	C	Uniform resampling	Tau-leaping, Tweedie denoiser analog	Geometric
DiscDiff Li et al. [2024]	D	DDPM	D	Uniform resampling	Absorb & Escape (A&E), Fast A&E	Not specified
MDLM Sahoo et al. [2024]	D	DDPM	D C	Absorbing	Efficient Ancestral Sampling, Semi-Autoregressive Sampling	Log Linear, Cosine Squared, Cosine, Linear

8.2 Discrete diffusion formulations

Denoising diffusion probabilistic models (DDPM) A denoising diffusion probabilistic model has two Markov chains: the forward chain that corrupts data and the backwards chain that recovers the data from the corrupted sample. The forward process is typically designed such that any data point converges to a simple prior distribution $\pi(x)$, typically a Gaussian.

The foundational work of Sohl-Dickstein et al. [2015] introduced the *diffusion probabilistic model*, which was later improved by Ho et al. [2020] into the well known *denoising diffusion probabilistic model* (DDPM). Following the first implementation of a fully discrete diffusion model by Forre et al. [2022], the work of Austin et al. [2021] generalized the DDPM formulation for discrete data. Their key innovation was the use of a transition matrix to define the forward process. This matrix can be explicitly designed to control how data is noised and to which prior distribution it converges. The framework proposed by Austin et al. [2021] is frequently used in subsequent work He et al. [2022]Lou et al. [2024]Sahoo et al. [2024], and is therefore briefly described here.

For a data sample $\boldsymbol{x}_0$ composed of one-hot encoded vectors of length K (representing K categories), the forward process is a Markov chain that generates a sequence of latent variables $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$. This process is defined by a series of transition matrices $\boldsymbol{Q}_t$, which dictate the iterative noising of the data. A single step in this forward process is defined as:

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) = \operatorname{Cat}\left(\boldsymbol{x}_t; \boldsymbol{p} = \boldsymbol{x}_{t-1}\boldsymbol{Q}_t\right), \tag{1}$$

where $\operatorname{Cat}(\boldsymbol{x}_t; \boldsymbol{p})$ is a categorical distribution. The probability vector $\boldsymbol{p}$ is the product of the one-hot row vector $\boldsymbol{x}_{t-1}$ and the transition matrix $\boldsymbol{Q}_t$.

Due to the chain rule of probability and the Markov property, the distribution of $\boldsymbol{x}_t$ at any timestep t given $\boldsymbol{x}_0$ can be computed in closed form by taking the cumulative product of the transition matrices $\overline{\boldsymbol{Q}}_t$:

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right) = \operatorname{Cat}\left(\boldsymbol{x}_t; \boldsymbol{p} = \boldsymbol{x}_0\overline{\boldsymbol{Q}}_t\right), \quad \text{with} \quad \overline{\boldsymbol{Q}}_t = \boldsymbol{Q}_1\boldsymbol{Q}_2 \ldots \boldsymbol{Q}_t \tag{2}$$

The forward process $q(x_t|x_{t-1})$ should be designed such that it's posterior $q(x_{t-1}|x_t, x_0)$ is tractable. Using Bayes' rule, it can be expressed as:

$$\begin{aligned} q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right) &= \frac{q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0\right) q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0\right)}{q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_0\right)} \\ &= \operatorname{Cat}\left(\boldsymbol{x}_{t-1}; \boldsymbol{p} = \frac{\boldsymbol{x}_t\boldsymbol{Q}_t^\top \odot \boldsymbol{x}_0\overline{\boldsymbol{Q}}_{t-1}}{\boldsymbol{x}_0\overline{\boldsymbol{Q}}_t\boldsymbol{x}_t^\top}\right). \end{aligned} \tag{3}$$

The model learns a reverse process, $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$, to approximate this true posterior. This is achieved by optimizing a variational upper bound on the negative log-likelihood, often referred to as the Evidence Lower Bound (ELBO):

$$
\begin{aligned}
L_{\text{vb}} = \mathbb{E}_{q(\boldsymbol{x}_0)}[& \underbrace{D_{\text{KL}}\left[q\left(\boldsymbol{x}_T \mid \boldsymbol{x}_0\right) \| p\left(\boldsymbol{x}_T\right)\right]}_{L_T} \\
& + \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)}\left[D_{\text{KL}}\left[q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right) \| p_\theta\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)\right]\right]}_{L_{t-1}} \\
& \underbrace{-\mathbb{E}_{q(\boldsymbol{x}_1 \mid \boldsymbol{x}_0)}\left[\log p_\theta\left(\boldsymbol{x}_0 \mid \boldsymbol{x}_1\right)\right]}_{L_0}].
\end{aligned} \tag{4}
$$

Each term in the objective has a distinct role; the term L_T is the difference between the distribution of a fully noised sample and our chosen prior $p(x_T)$. When the forward process is designed to converge to a known stationary distribution when t reaches T regardless of x_0, this term becomes zero. The L_{t-1} terms drive the learning process by minimizing the KL divergence between the learned reverse step $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ and the true posterior $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$. Finally, the L_0 term accounts for the final timestep that bridges the slightly noisy x_1 to the uncorrupted data $\boldsymbol{x}_0$. By minimizing this negative log-likelihood term, the training objective forces the model to assign high probability to the true data x_0 when conditioned on its slightly noisy counterpart x_1. It effectively serves as the loss for the final step of the reverse diffusion trajectory, ensuring that the model can accurately map the slightly noisy latent state back to the clean data manifold.

Score-based generative models (SGMs) Unlike likelihood-based models that optimise probabilities directly, score-based generative models rely on the score function Hyvarinen [2005]. For a probability density function $p(x) = \frac{e^{-f_\theta(x)}}{Z_\theta}$, the score function is given by the gradient of its log-density $\nabla_x \log p(x)$. Intuitively, it can be interpreted as a vector field pointing in the direction of the steepest increase in probability density.

The formulation of the score function, $\nabla_x \log p_\theta(x)$, is a key advantage of score-based models because it allows for a derivation where the often intractable normalising constant Z_θ is eliminated Hyvarinen [2005], Song and Ermon [2020]. The derivation proceeds as follows:

$$
\begin{aligned}
\nabla_x \log p_\theta(x) &= \nabla_x \log \frac{e^{-f_\theta(x)}}{Z_\theta} \\
&= \nabla_x \left(\log e^{-f_\theta(x)} - \log Z_\theta\right) && \text{(Logarithm of a quotient)} \\
&= \nabla_x \log e^{-f_\theta(x)} - \nabla_x \log Z_\theta && \text{(Linearity of the gradient)} \\
&= \nabla_x(-f_\theta(x)) - \nabla_x \log Z_\theta && \text{(Logarithm and exponential cancel)} \\
&= -\nabla_x f_\theta(x) - 0 && \text{(Since } Z_\theta \text{ is not a function of } x\text{)} \\
&= -\nabla_x f_\theta(x)
\end{aligned} \tag{5}
$$

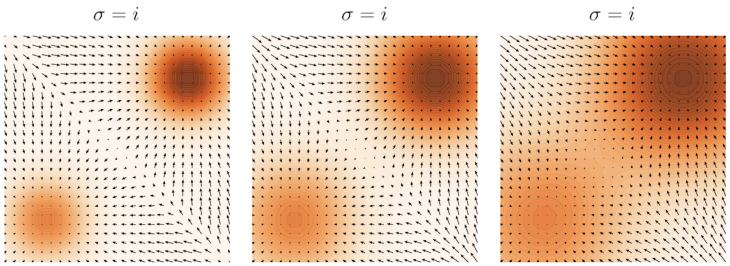

Fig. 3. Illustration of a score function with increasing levels of noise added to the data. The arrows represent the score function, a vector field that points in the direction of increasing data probability. The orange colormap indicates the probability density of the data. > (Left) At the start of the process (low noise), the score function can be ill-defined in regions of low data density. > (Center to Right) As noise is added over time, the data distribution diffuses and becomes smoother. This allows the model to learn a stable score function that is well-behaved across the entire data space. This Figure is inspired by Figure 2 from Song and Ermon [2020]

.

In this derivation, the term $\nabla_x \log Z_\theta$ vanishes. This is because the normalisation constant $Z_\theta = \int e^{-f_\theta(x')} dx'$ is a function of the model parameters θ and does not depend on any specific value of x. Consequently, $\log Z_\theta$ is also a constant with respect to x, and its gradient $\nabla_x \log Z_\theta$ is zero. This means the score function can be computed and learned without needing to evaluate Z_θ, which is a major advantage for models where Z_θ is intractable Song and Ermon [2020].

However learning a score model is not trivial and the output of score is frequently incorrect for positions that lay outside of the data distribution that the model was trained on. To perform better in these low data density locations, score-based generative models typically perturb the input data with Gaussian noise at various levels, and then learn a score-based model conditioned on the noise level (forming a Noise Conditional Score Network, or NCSN) Song and Ermon [2020]. This process of adding noise diffuses the original data distribution (as illustrated in Figure 4), causes regions that were initially of low data density in the unperturbed space to be more frequently encountered as noisy samples during training.

The score function $\nabla_x \log p_\theta(x)$ is only applicable to continuous space Vincent [2011]Meng et al. [2023]. Many adaptations of

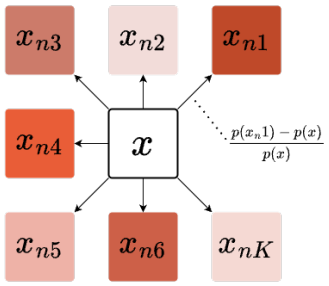

Fig. 4. Visualisation of the *Concrete Score* for a data point x and its local neighborhood $\mathcal{N}(x) = x_{n1}, ..., x_{nk}$. The color intensity of each neighbor x_{ni} indicates the relative change in probability for the transition from x, with darker colors signifying a more favorable transition to a higher probability state.

a score function that work in discrete space have been considered Hyvärinen [2007]Hyvarinen [2007]Lyu [2012]] however the one that has been adopted within the field is the *Concrete Score* Meng et al. [2023]. The concrete score works by considering a set of neighbours $\mathcal{N}(x) = x_{n1}, ..., x_{nk}$ for a datapoint x. The concrete score $\boldsymbol{c}_{p_{\text{data}}}(\mathbf{x}; \mathcal{N})$ for a given data distribution $p_{\text{data}}(x)$ is:

$$\boldsymbol{c}_{p_{\text{data}}}(\mathbf{x}; \mathcal{N}) \triangleq \left[\frac{p_{\text{data}}(\mathbf{x}_{n_1}) - p_{\text{data}}(\mathbf{x})}{p_{\text{data}}(\mathbf{x})}, \ldots, \frac{p_{\text{data}}(\mathbf{x}_{n_k}) - p_{\text{data}}(\mathbf{x})}{p_{\text{data}}(\mathbf{x})} \right]^T \tag{6}$$

The term $\frac{p(x_{n_i}) - p(x)}{p(x)}$ represents the relative change in probability when moving from x to x_{n_i}, and is evaluated for each neighbor (x_{n_i}). The neighborhood structure $\mathcal{N}(x)$ defines a directed graph over the data points, and the Concrete Score provides values for each edge (x, x_{n_i}) indicating the local attractiveness of transitioning to that neighbor in terms of probability. Meng et al. [2023] also defined concrete score matching, which is the objective for a concrete score model $c_\theta(\cdot; \mathcal{N})$:

$$\mathcal{L}_{\text{CSM}}(\theta) = \sum_{\mathbf{x}} p_{\text{data}}(\mathbf{x}) \left\| \boldsymbol{c}_\theta(\mathbf{x}; \mathcal{N}) - \boldsymbol{c}_{p_{\text{data}}}(\mathbf{x}; \mathcal{N}) \right\|_2^2 \tag{7}$$

where $c_\theta(x; \mathcal{N})$ denotes the score model and $c_{p_{\text{data}}}(x; \mathcal{N})$ the true score. the objective function is then simply the average squared Euclidean distance between the learned score and the true score. However, this objective is intractable as it requires the true score $\boldsymbol{c}_{p_{\text{data}}}$, which depends on the unknown data distribution. To overcome this, Meng et al. [2023] derived a practical, tractable objective that avoids this dependency. For the full derivation, we refer the reader to the original work.

Continuous time frameworks The DDPM framework discussed above works with a discrete time variable, where t takes integer values from a set $\{1, \ldots, T\}$. In this formulation, the diffusion process is a discrete-time Markov chain, where each step can be seen as a generalization of a Bernoulli process to a categorical draw. However, the time variable can also be treated as continuous. For diffusion models operating on continuous data, this has been done by formulating the forward and backward diffusion processes as stochastic differential equations (SDEs) Song et al. [2020]. This SDE-based approach was shown to generate higher-quality samples, achieving remarkable results on image generation benchmarks. Beyond improved sample quality, a key advantage of this continuous-time perspective is the derivation of an equivalent ordinary differential equation (ODE) from the reverse-time SDE. This deterministic ODE formulation is particularly beneficial as it enables exact log-likelihood computation and more flexible, faster sampling algorithms Song et al. [2020].

The analog for extending diffusion to discrete data spaces is formulating the process as a Continuous-Time Markov Chain (CTMC). This continuous-time formulation unlocks several significant advantages over its discrete-time counterpart. A key motivation is the enhanced flexibility it affords; because the model learns a process defined for any continuous time t, it is not constrained to a fixed number of denoising steps. This allows for the use of more sophisticated and efficient sampling algorithms, such as adaptive samplers, which can improve generation quality and computational efficiency Campbell et al. [2022], Liu et al. [2024]. From a theoretical standpoint, the CTMC framework leads to simpler and more elegant training objectives, enabling simplifying the Evidence Lower Bound (ELBO) to a weighted integral of cross-entropy losses Sahoo et al. [2024], Shi et al. [2025]. Crucially, these benefits are not merely theoretical; ablation studies have empirically demonstrated that continuous-time formulations can marginally outperform their discrete-time counterparts in terms of perplexity Sahoo et al. [2024].

A continuous-time framework for diffusion models in discrete state spaces was first explored theoretically by Austin et al. [2021]. The work of Campbell et al. [2022] further developed and implemented this framework, providing a practical foundation that many subsequent models have built upon. We will now give a brief overview of this framework.

A CTMC can be conceptualized as a process where a variable transitions between discrete states at random moments in continuous time. Its evolution is governed by the Markov property, meaning the future state depends only on the current state. The process of transitioning is governed by the transition rate matrix, R_t, which is formally defined as:

$$R(\tilde{x}, x) = \lim_{\Delta t \to 0} \frac{q_{t|t-\Delta t}(x \mid \tilde{x}) - \delta_{x,\tilde{x}}}{\Delta t} \tag{8}$$

where $R(\tilde{x}, x)$ is an element of the rate matrix, and $q_{t|t-\Delta t}(x|\tilde{x})$ is the infinitesimal transition probability of being in state x at time t given the process was in state $\tilde{x}$ at time $t - \Delta t$. The Kronecker delta, $\delta_{x,\tilde{x}}$, is 1 if $x = \tilde{x}$ and 0 otherwise. Its role is to separate the calculation for transitions between different states from the calculation for remaining in the same state. Consequently, the off-diagonal elements of R_t hold the non-negative instantaneous rates of transitioning to a neighboring state, while the diagonal elements hold the non-positive rates of leaving the current state, ensuring that the rows of the matrix sum to zero.

Conversely, the transition probability for an infinitesimally small time step Δt can be expressed as:

$$q_{t|t-\Delta t}(x \mid \tilde{x}) = \delta_{x,\tilde{x}} + R(\tilde{x}, x)\Delta t + o(\Delta t) \tag{9}$$

The transition rate matrix R_t can be viewed as the continuous-time analog of the transition probability matrix Q_t from the discrete DDPM framework Austin et al. [2021]. It can also be designed to enable different forward processes, such as uniform or absorbing diffusion.

Calculating the transition probabilities over a finite interval requires integrating the Kolmogorov differential equations Campbell et al. [2022]. This can be done analytically if the rate matrices at different times commute (i.e., R_t and $R_{t'}$ commute for all t, t'). A practical way to satisfy this condition is to define the rate matrix as $R_t = \beta(t)R_b$, where R_b is a time-independent base rate matrix defining the structure of transitions (e.g., uniform), and $\beta(t)$ is a scalar noise schedule that controls the speed of noising over time. The authors of Campbell et al. [2022] provide the analytical solution for the forward process transition probabilities, $q_{t|0}(x = j|x_0 = i)$, which is obtained by solving the Kolmogorov forward equation for their specific choice of rate matrix R_t.

8.3 Search strategy

Papers included in this survey were found by initially searching *Google scholar* and *Ar-xiv* with the search term: "Discrete diffusion", "Discrete diffusion Natural Language", "Discrete diffusion DNA". Additionally papers mentioned in related works sections of papers were also added to the list, recursively, until no new papers were found.