Flatten Graphs as Sequences: Transformers are Scalable Graph Generators

Dexiong Chen, Markus Krimmel, Karsten Borgwardt

Max Planck Institute of Biochemistry

NeurIPS 2025

MAX PLANCK INSTITUTE
OF BIOCHEMISTRY

MLSB

Representing graphs as sequences

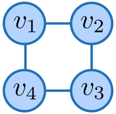
What is the “language” of a graph?

Representing graphs as sequences

What is the "language" of a graph? $\Rightarrow$ Random walks!

Representing graphs as sequences

What is the “language” of a graph? $\Rightarrow$ Random walks!

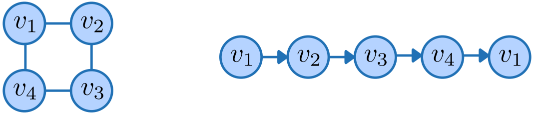


Eulerian trail

An Eulerian trail (or Eulerian path) is a trail in a finite graph that visits every edge exactly once.

Representing graphs as sequences

What is the “language” of a graph? $\Rightarrow$ Random walks!



Eulerian trail

An Eulerian trail (or Eulerian path) is a trail in a finite graph that visits every edge exactly once.

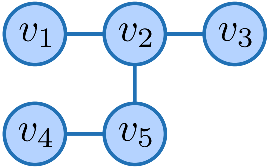
But most graphs don’t have Eulerian trails according to Euler’s theorem...

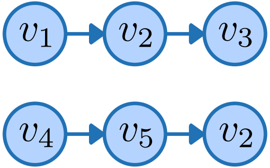
Extension of Eulerian trail

Segmented Eulerian trail (SET)

A segmented Eulerian trail (SET) is a sequence of trail segments such that each edge is visited exactly once across all segments, and segments do not need to be connected.

Extension of Eulerian trail





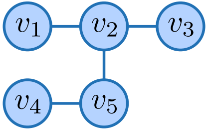
Segmented Eulerian trail (SET)

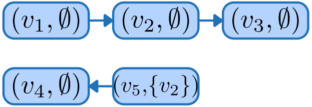
A segmented Eulerian trail (SET) is a sequence of trail segments such that each edge is visited exactly once across all segments, and segments do not need to be connected.

Remarks:

- Each graph has multiple SETs, but each SET can induce only one graph.
- The graph defined by any prefix of a SET is a subgraph of the original graph, but not necessarily an induced subgraph.

Extending SET to incorporate neighborhood information

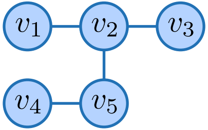


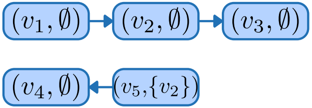


Segmented Eulerian neighborhood trail (SENT)

- Each node in a SET will be paired with a neighborhood set which includes all visited nodes that are neighbors to this node.
- Each edge is still visited exactly once across all trails and neighborhood sets.

Extending SET to incorporate neighborhood information





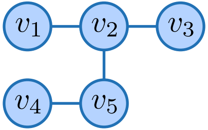
Segmented Eulerian neighborhood trail (SENT)

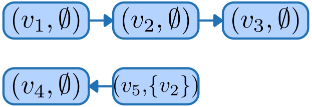
- Each node in a SET will be paired with a neighborhood set which includes all visited nodes that are neighbors to this node.
- Each edge is still visited exactly once across all trails and neighborhood sets.

Key properties:

- Prefixes of any SENT are induced subgraphs of the original graph.

Extending SET to incorporate neighborhood information





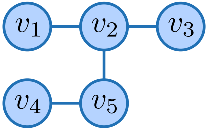
Segmented Eulerian neighborhood trail (SENT)

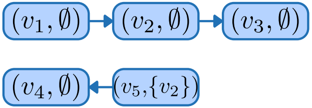
- Each node in a SET will be paired with a neighborhood set which includes all visited nodes that are neighbors to this node.
- Each edge is still visited exactly once across all trails and neighborhood sets.

Key properties:

- Prefixes of any SENT are induced subgraphs of the original graph.
  $\Rightarrow$ Substructure-conditioned generation can be achieved for free!

Extending SET to incorporate neighborhood information





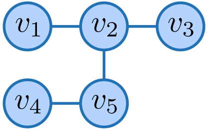
Segmented Eulerian neighborhood trail (SENT)

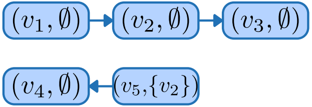
- Each node in a SET will be paired with a neighborhood set which includes all visited nodes that are neighbors to this node.
- Each edge is still visited exactly once across all trails and neighborhood sets.

Key properties:

- Prefixes of any SENT are induced subgraphs of the original graph.
  $\Rightarrow$ Substructure-conditioned generation can be achieved for free!
- Each node is visited once in the trails (but not in the neighborhood sets).

Extending SET to incorporate neighborhood information





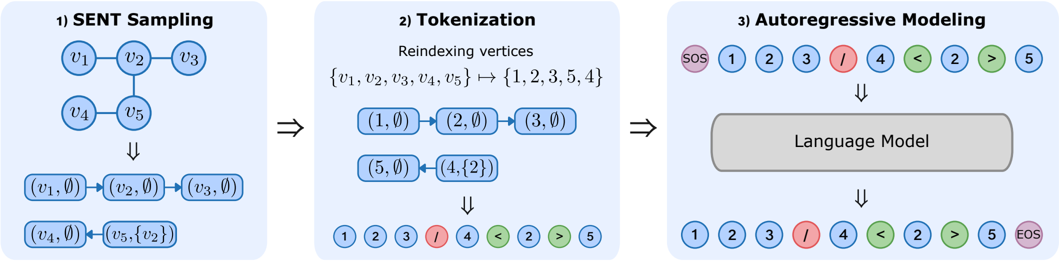
Segmented Eulerian neighborhood trail (SENT)

- Each node in a SET will be paired with a neighborhood set which includes all visited nodes that are neighbors to this node.
- Each edge is still visited exactly once across all trails and neighborhood sets.

Key properties:

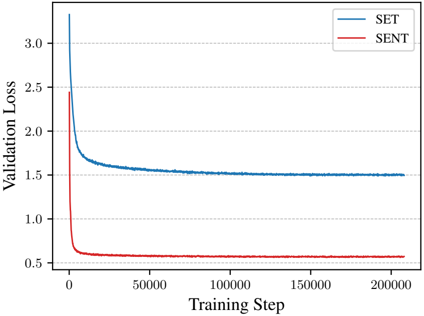
- Prefixes of any SENT are induced subgraphs of the original graph.
  $\Rightarrow$ Substructure-conditioned generation can be achieved for free!
- Each node is visited once in the trails (but not in the neighborhood sets).
  $\Rightarrow$ A SENT can be sampled efficiently via random path sampling.

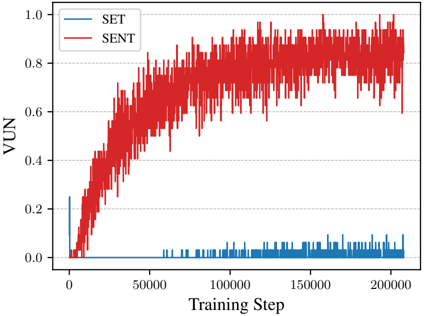
AutoGraph: graph generation as language modeling



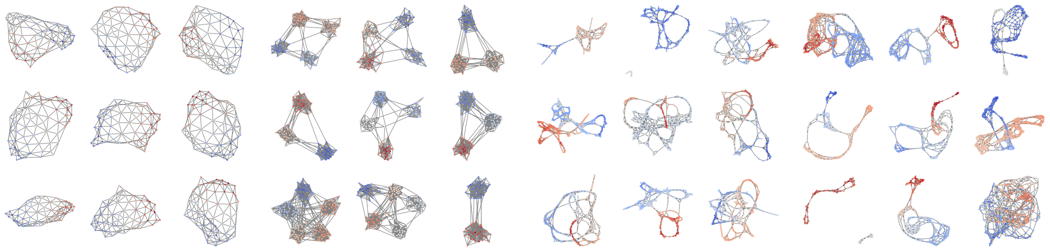
- Vertices reindexing relies on the node ordering in the SENT.
- Special tokens are used to indicate breaks between segments, and the start and end of neighborhood sets.
- Graph generation is recast as a potentially easier sequence generation problem.

Experiment: SET vs SENT





Results on synthetic graph datasets



Planar Graphs $n_{\text{graphs}} = 128$, $\lvert V \rvert = 64$		
Model	MMD$_R$↓	VUN↑
SOTA	1.8	90.0
AutoGraph	1.5	87.5

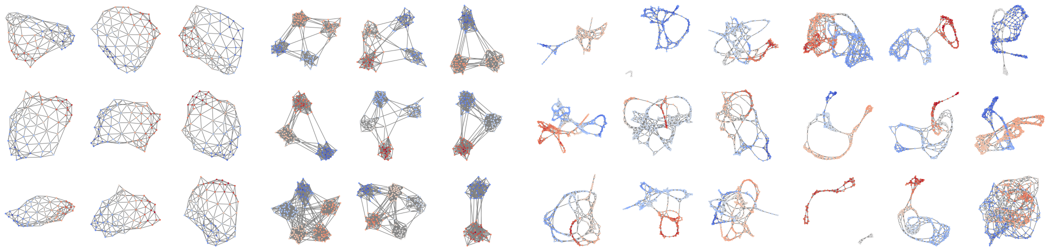
Stochastic Block Models $n_{\text{graphs}} = 128$, $\lvert V \rvert_{\text{avg}} \approx 104$		
Model	MMD$_R$↓	VUN↑
SOTA	1.5	85.0
AutoGraph	3.4	92.5

Proteins $n_{\text{graphs}} = 587$, $\lvert V \rvert_{\text{avg}} \approx 258$		
Model	MMD$_R$↓	VUN↑
SOTA	4.7	–
AutoGraph	2.3	–

Point Clouds $n_{\text{graphs}} = 26$, $\lvert V \rvert_{\text{avg}} \approx 1332$		
Model	MMD$_R$↓	VUN↑
SOTA	6.8	–
AutoGraph	3.0	–

Vignac et al. [2022], Bergmeister et al. [2024], Jo et al. [2024]

Results on synthetic graph datasets



Planar Graphs $n_{\text{graphs}} = 128$, $\lvert V\rvert = 64$		
Model	MMD$_R$↓	VUN↑
SOTA	1.8	90.0
AutoGraph	1.5	87.5

Stochastic Block Models $n_{\text{graphs}} = 128$, $\lvert V\rvert_{\text{avg}} \approx 104$		
Model	MMD$_R$↓	VUN↑
SOTA	1.5	85.0
AutoGraph	3.4	92.5

Proteins $n_{\text{graphs}} = 587$, $\lvert V\rvert_{\text{avg}} \approx 258$		
Model	MMD$_R$↓	VUN↑
SOTA	4.7	–
AutoGraph	2.3	–

Point Clouds $n_{\text{graphs}} = 26$, $\lvert V\rvert_{\text{avg}} \approx 1332$		
Model	MMD$_R$↓	VUN↑
SOTA	6.8	–
AutoGraph	3.0	–

AutoGraph is 100x faster than discrete diffusion models (DiGress) in sampling!

Vignac et al. [2022], Bergmeister et al. [2024], Jo et al. [2024]

Extension to attributed graphs

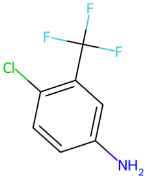
AutoGraph can be extended to graphs with node and edge attributes

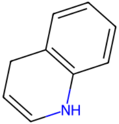
$\Rightarrow$ by inserting the attributes into the SENT sequence in an interleaved fashion.

	GuacaMol				
	$n_{\text{graphs}} = 1.1$M, $\lvert V\rvert_{\text{avg}} \approx 28$				
Model	Valid↑	Unique↑	Novel↑	KL div↑	FCD↑
SOTA	85.2	100	99.9	92.9	68.0
AutoGraph	91.6	100	97.7	97.5	79.2
AutoGraph (pretrained)	95.9	100	95.5	98.1	91.4

Unconditional generation

Conditional generation





Conditional generation

Take-home messages

- A powerful, scalable, and flexible model for attributed graph generation
- Bridge the gap between language modeling and graph generation
- Potential applications for drug discovery, protein design, etc
- Towards graph foundation models for biology

Take-home messages

- A powerful, scalable, and flexible model for attributed graph generation
- Bridge the gap between language modeling and graph generation
- Potential applications for drug discovery, protein design, etc
- Towards graph foundation models for biology

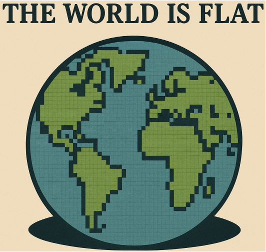


Image created by ChatGPT

References I

A. Bergmeister, K. Martinkus, N. Perraudin, and R. Wattenhofer. Efficient and scalable graph generation through iterative local expansion. In *International Conference on Learning Representations (ICLR)*, 2024.

J. Jo, D. Kim, and S. J. Hwang. Graph generation with diffusion mixture. In *International Conference on Machine Learning (ICML)*. PMLR, 2024.

C. Vignac, I. Krawczuk, A. Siraudin, B. Wang, V. Cevher, and P. Frossard. Digress: Discrete denoising diffusion for graph generation. In *International Conference on Learning Representations (ICLR)*, 2022.