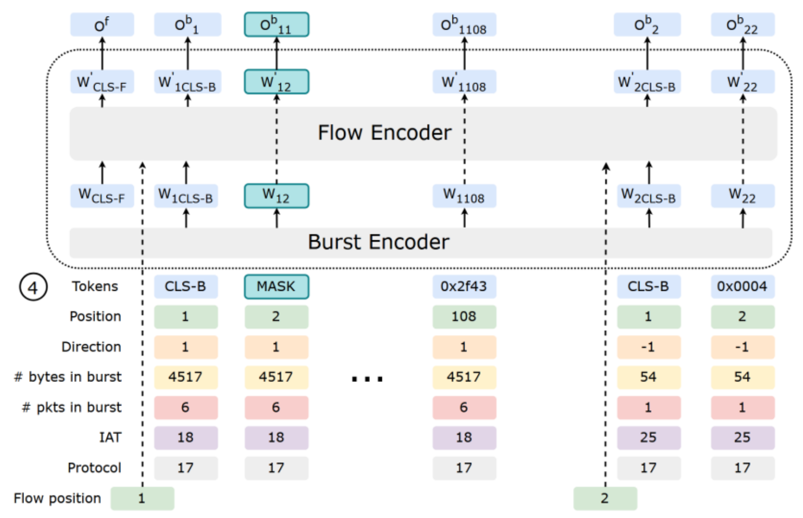
Demystifying Network Foundation Models

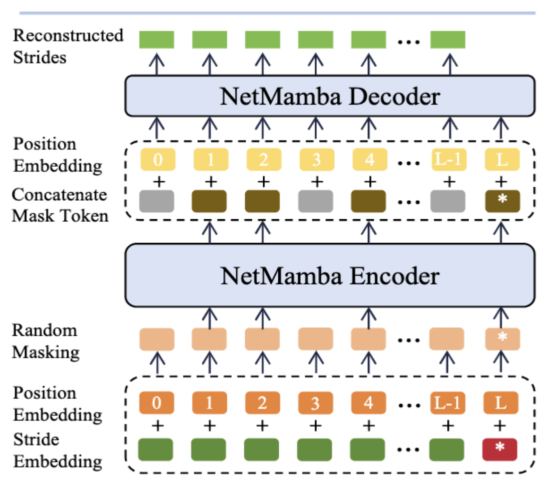
Sylee (Roman) Beltiukov[1], Satyandra Guthula[1], Wenbo Guo[1], Walter Willinger[2], Arpit Gupta[1]

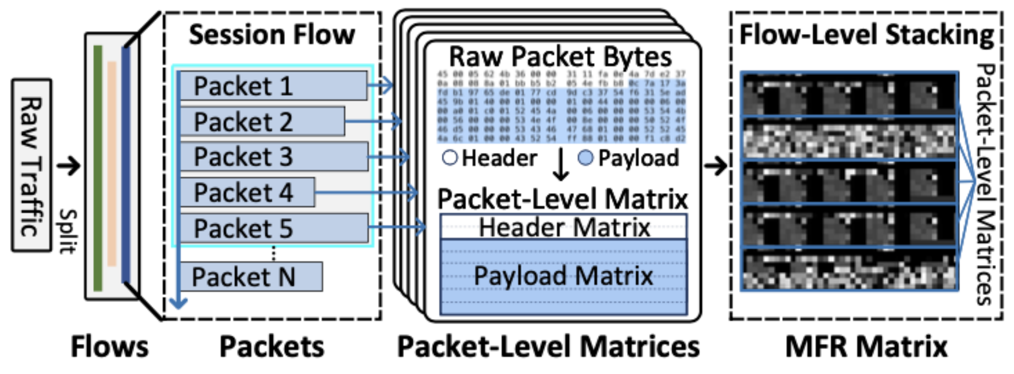
[1]UC Santa Barbara, [2]NIKSUN

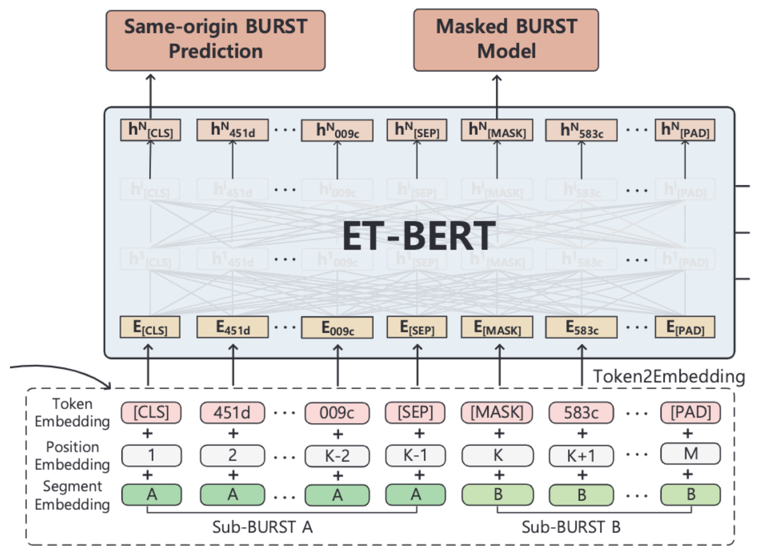
Foundation Models in Networking!

- Network Foundation Models (NFMs) as an answer to generalizability over various infrastructures
- Leveraging massive amount of unlabeled network data



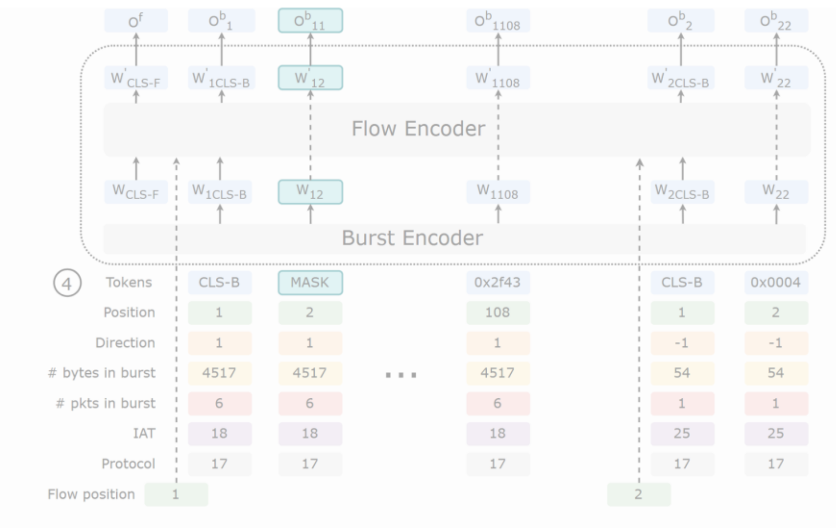


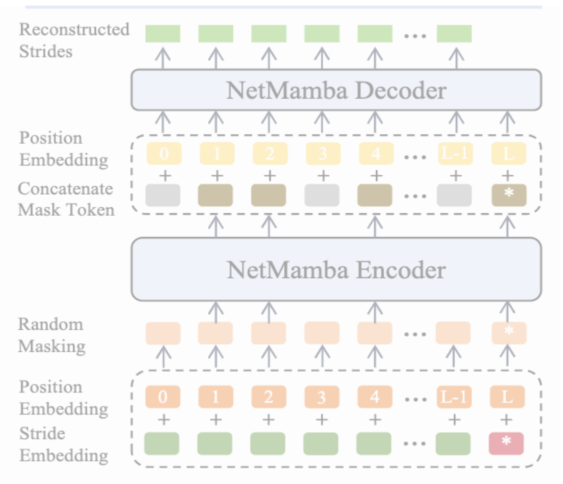


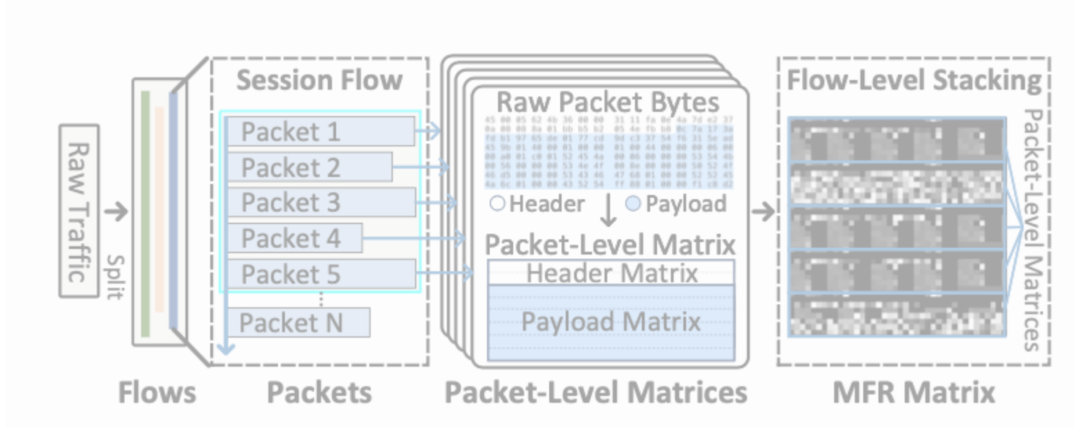


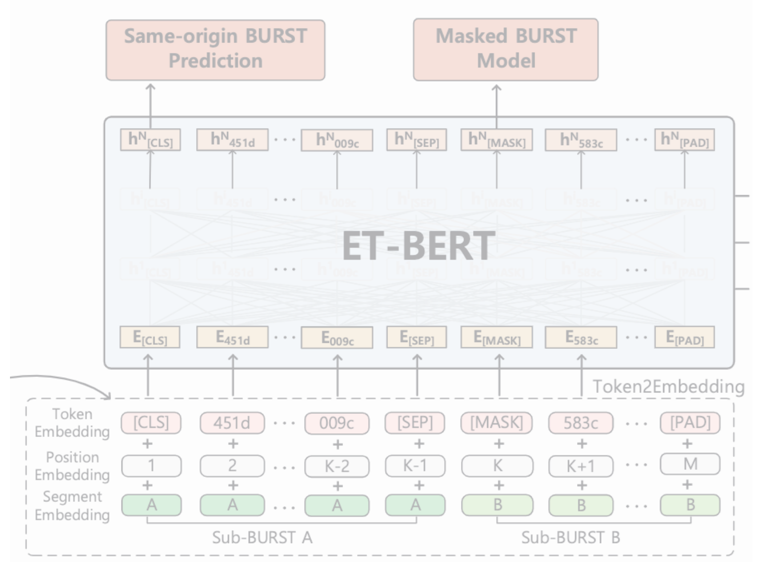
Foundation Models in Networking!

- Network Foundation Models (NFMs) as an answer to generalizability over various infrastructures
- Leveraging massive amount of unlabeled network data









- High expectations & excitement
- Wonderful numbers in every benchmark paper

	Model #1	Model #2	…
Fine-tuning dataset #1	99.95	99.96	…
Fine-tuning dataset #2	99.54	99.55	…
…	…	…	…

But do we evaluate their knowledge correctly?

*

	CIC-IDS (Heartbleed)		**Crossmarket** (*Acc*@10)	
	Original	Fixed	Original	Fixed
ET-BERT	99.99 ± 0.01		99.82 ± 0.03	
YaTC	99.99 ± 0.01		99.69 ± 0.03	

*Guthula et al. netFound: Foundation Model for Network Security. *arXiv: 2310.17025*

But do we evaluate their knowledge correctly?

*

	CIC-IDS (Heartbleed)		Crossmarket ($Acc@10$)	
	Original	Fixed	Original	Fixed
ET-BERT	99.99 ± 0.01	0.0 ± 0.0	99.82 ± 0.03	35.62 ± 0.39
YaTC	99.99 ± 0.01	0.01 ± 0.01	99.69 ± 0.03	58.13 ± 0.89

*Guthula et al. netFound: Foundation Model for Network Security. *arXiv: 2310.17025*

But do we evaluate their knowledge correctly?

*

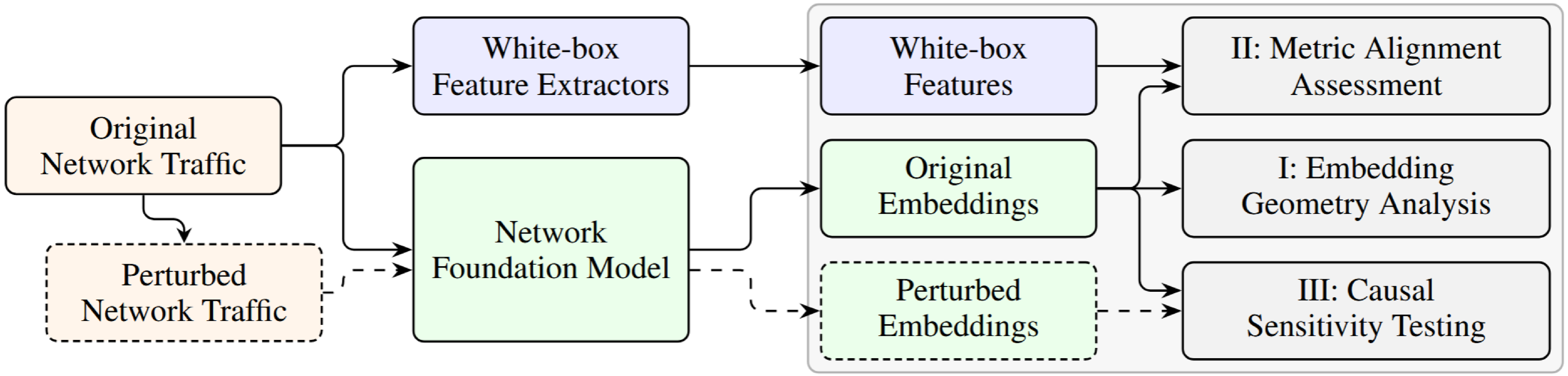
	CIC-IDS (Heartbleed)		Crossmarket (*Acc@10*)	
	Original	Fixed	Original	Fixed
ET-BERT	99.99 ± 0.01	0.0 ± 0.0	99.82 ± 0.03	35.62 ± 0.39
YaTC	99.99 ± 0.01	0.01 ± 0.01	99.69 ± 0.03	58.13 ± 0.89

- Equating NFMs' success with performance on a limited set of downstream tasks and datasets is misleading
- What exactly pretraining gives?
- Can we somehow *uncover* what latent knowledge is inside pretrained models?

*Guthula et al. netFound: Foundation Model for Network Security. *arXiv: 2310.17025*

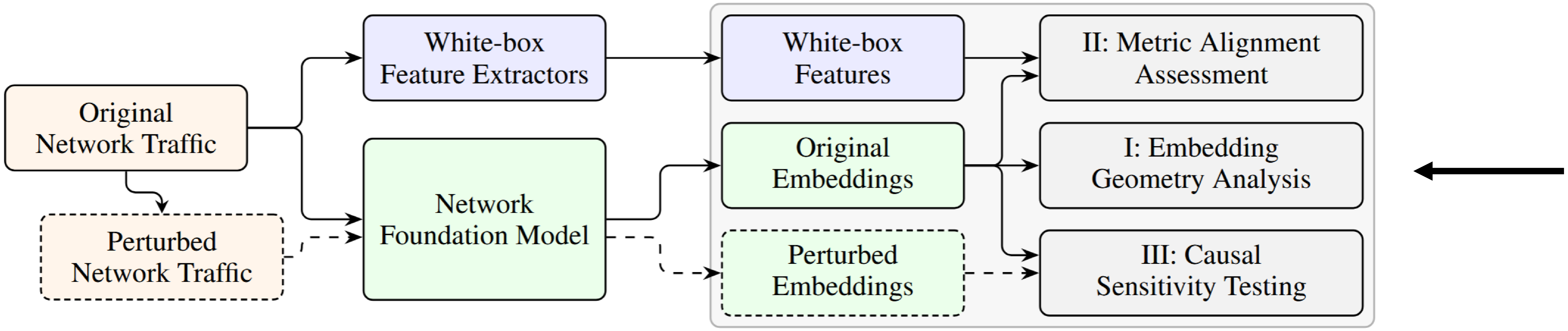
Intrinsic Evaluation Framework

Assess embedding quality decoupled from downstream tasks/datasets



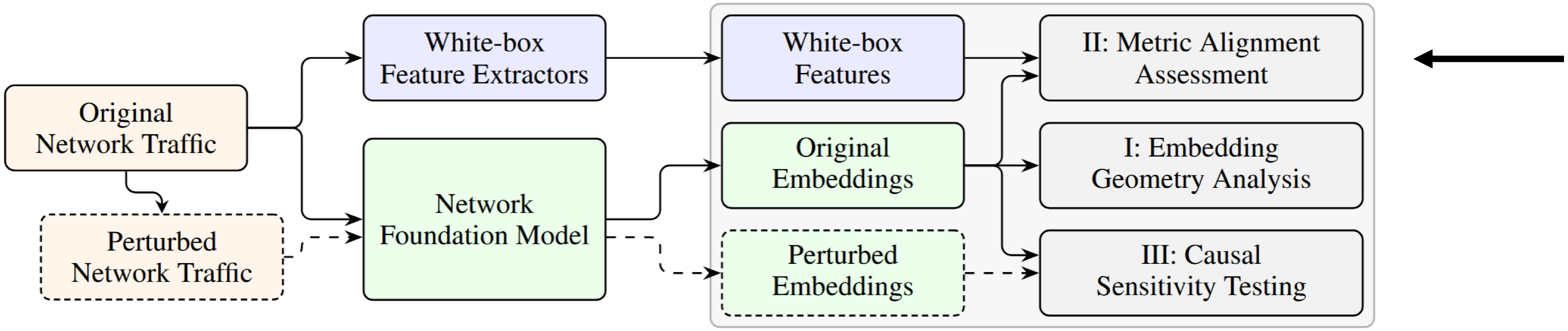
Intrinsic Evaluation Framework

Assess embedding quality decoupled from downstream tasks/datasets



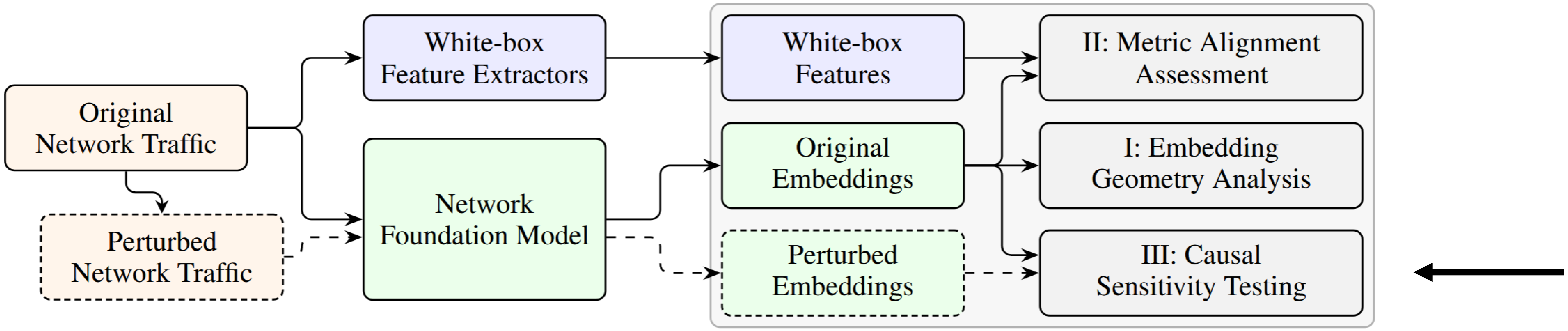
Intrinsic Evaluation Framework

Assess embedding quality decoupled from downstream tasks/datasets



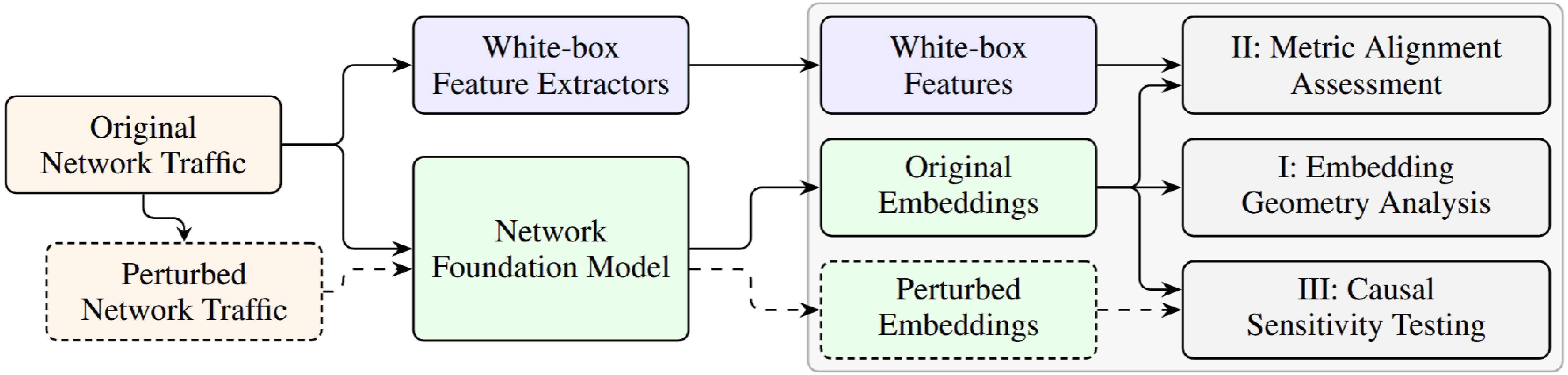
Intrinsic Evaluation Framework

Assess embedding quality decoupled from downstream tasks/datasets



Intrinsic Evaluation Framework

Assess embedding quality decoupled from downstream tasks/datasets



For evaluation, we took:

- All NFMs with available code and weights[1]

  YaTC, ET-BERT, netFound, NetMamba

- Several endogenous and exogenous datasets

  Android Crossmarket, CIC-IDS17, CIC-APT-IIoT24
  MAWI, CAIDA

[1]*Spring 2025*

Embedding Geometry Analysis

Idea: embeddings should fully utilize representation space instead of clustering together

- We measured random pairwise cosine similarity (anisotropy)[1]
  Including Mean Cosine Contribution of each dimension

[1]Kawin Ethayarajh. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings.
In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing

Embedding Geometry Analysis

Idea: embeddings should fully utilize representation space instead of clustering together

- We measured random pairwise cosine similarity (anisotropy)[1]

  Including Mean Cosine Contribution of each dimension

- It allowed us to uncover different failure modes…

	Avg cos	*Top MCC*
NetMamba	0.96	0.02
YaTC	0.86	0.23
…	…	…

← Highly clusterized embeddings

[1]Kawin Ethayarajh. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*

Embedding Geometry Analysis

Idea: embeddings should fully utilize representation space instead of clustering together

- We measured random pairwise cosine similarity (anisotropy)[1]

  Including Mean Cosine Contribution of each dimension

- It allowed us to uncover different failure modes…

	Avg cos	*Top MCC*
NetMamba	0.96	0.02
YaTC	0.86	0.23
…	…	…

One dimension is overused

[1]Kawin Ethayarajh. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*

Embedding Geometry Analysis

Idea: embeddings should fully utilize representation space instead of clustering together

- We measured random pairwise cosine similarity (anisotropy)[1]
  Including Mean Cosine Contribution of each dimension
- It allowed us to uncover different failure modes...

	Avg cos	*Top MCC*
NetMamba	0.96	0.02
YaTC	0.86	0.23
...	...	...

And even improve F_1 score significantly with just whitening!

	Crossmarket	CIC-IDS2017	CIC-APT-IIoT24
NetMamba	+0.35±0.02	+0.11±0.27	+0.03±0.02
...	...	...	...

[1]Kawin Ethayarajh. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*

Metric Alignment Assessment

Idea: embeddings should be aware of well-known whitebox network features

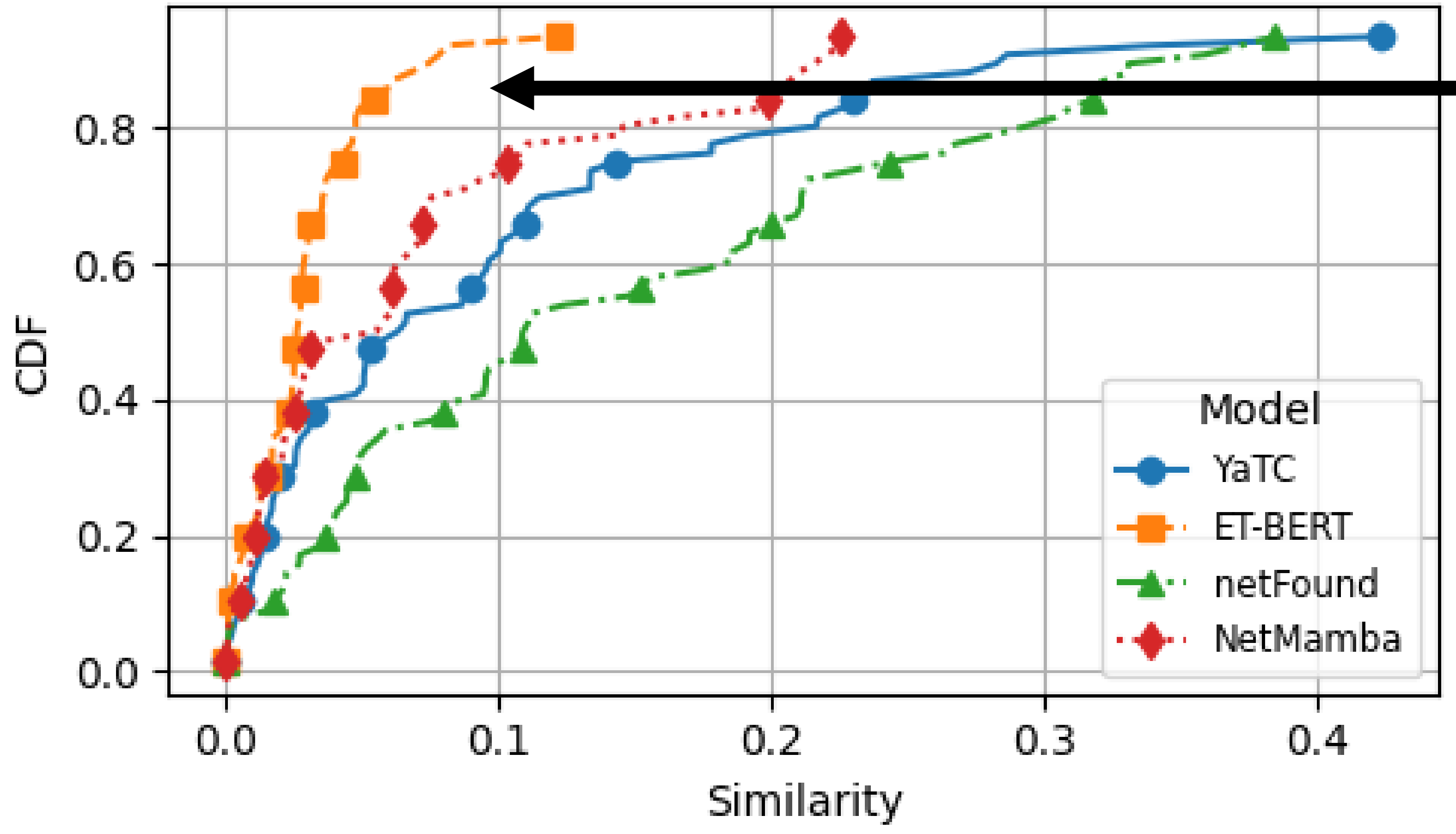
- We extracted meaningful statistical features from network traffic[1] and observed similarity between them and embeddings

[1]Using CICFlowMeter and Centered Kernel Alignment similarity index

Metric Alignment Assessment

Idea: embeddings should be aware of well-known whitebox network features

- We extracted meaningful statistical features from network traffic[1] and observed similarity between them and embeddings



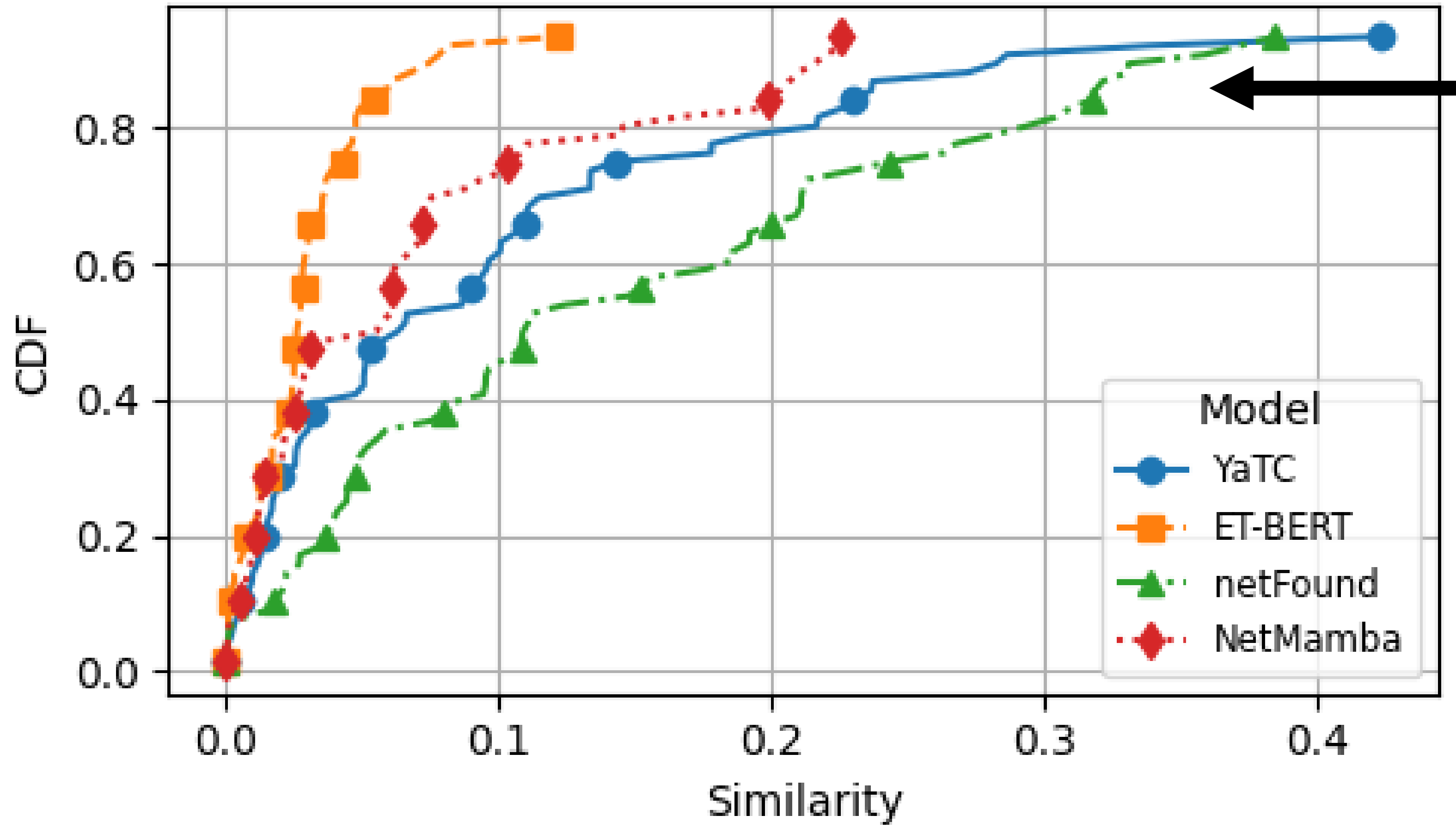
Payload-only models do not capture vital network statistics

[1]Using CICFlowMeter and Centered Kernel Alignment similarity index

Metric Alignment Assessment

Idea: embeddings should be aware of well-known whitebox network features

- We extracted meaningful statistical features from network traffic[1] and observed similarity between them and embeddings



Feature diversity enhances metric coverage

[1]Using CICFlowMeter and Centered Kernel Alignment similarity index

Metric Alignment Assessment

Idea: embeddings should be aware of well-known whitebox network features

- We extracted meaningful statistical features from network traffic[1] and observed similarity between them and embeddings

But <u>all</u> models still struggle significantly with real-world datasets!

Average feature similarity	
Real-world datasets (CAIDA, MAWI)	Controlled environments (CIC-IDS17, CIC-APT-IIoT24)
0.044	0.111

[1]Using CICFlowMeter and Centered Kernel Alignment similarity index

Causal Sensitivity Testing

Idea: embeddings should vary depending on importance of perturbations

- We applied various input data perturbations and observed output embeddings' changes

Causal Sensitivity Testing

Idea: embeddings should vary depending on importance of perturbations

- We applied various input data perturbations and observed output embeddings' changes

	YaTC		**ET-BERT**		**netFound**		**NetMamba**	
	% tok	*cos*	% tok	*cos*	% tok	*cos*	% tok	*cos*
...	...	...	...	...	...	...	...	...
Payload	75%		100%		33%		75%	

Causal Sensitivity Testing

Idea: embeddings should vary depending on importance of perturbations

- We applied various input data perturbations and observed output embeddings' changes

	YaTC		**ET-BERT**		**netFound**		**NetMamba**	
	% tok	*cos*	% tok	*cos*	% tok	*cos*	% tok	*cos*
...	...	...	...	...	...	...	...	...
Payload	75%	$\boxed{0.18}$	100%	$\boxed{0.48}$	33%	0.99	75%	$\boxed{0.62}$

- Embeddings vary a lot as response to encrypted payload?!
- Models rely on payload memorization!
  - They definitely shouldn't 👀

There’s more to demystify and understand

- Our framework provides insights into NFMs performance

  By investigating pretrained models’ quality

- But there is much more to learn and uncover

- And our understanding of NFMs remains limited

- See more examples and bold statements in the paper

  Example: some NFMs can observe fluctuations in AQM, Congestion Control, and even cross-traffic! 👀

- Code: https://github.com/maybe-hello-world/demystifying-networks

  Fully reproducible and ready to evaluate your network foundation model