Exploring End-to-end Differentiable Neural Charged Particle Tracking – A Loss Landscape Perspective – [1]

Tobias Kortus, Ralf Keidel, Nicolas R. Gauger
on behalf of the *Bergen pCT collaboration*

Scientific Comuting Group
NHR @ SW - Method Lab AI and ML (https://nhrsw.de/mls/)
University of Kaiserslautern-Landau (RPTU)

4th MODE Workshop on Differentiable Programming for Experiment Design

Valencia, September 23, 2024

NHR@SW

[1] Based on: Tobias Kortus, Ralf Keidel, & Nicolas R. Gauger. (2024). Exploring End-to-end Differentiable Neural Charged Particle Tracking – A Loss Landscape Perspective. https://arxiv.org/abs/2407.13420

Motivation: Particle Tracking

- **Bergen pCT detector:** Multilayer sandwich tracking calorimeter developed for the application of proton computed tomography (pCT).
  - 2 tracking- & 41 detector-absorber layer
  - Measurement of path & residual energy after patient, used for image reconstruction.
- **Goal**: Reconstruct path of protons under the influence of elastic and inelastic interactions (mostly scattering).

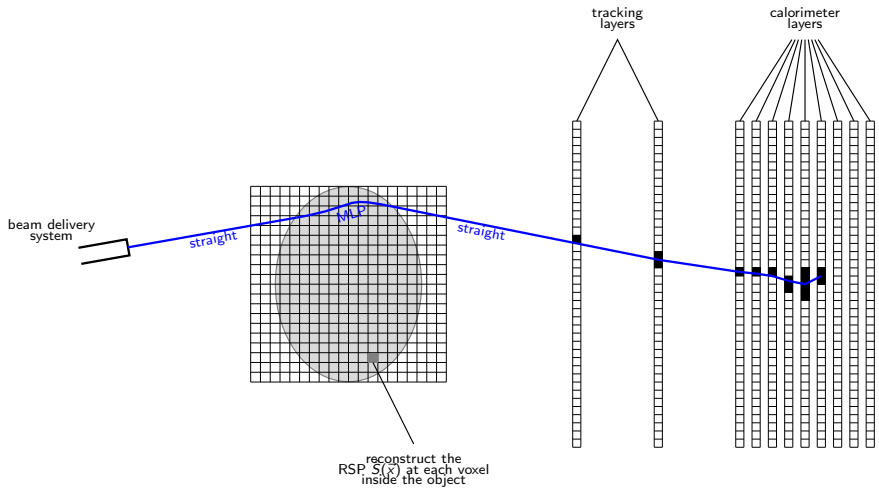
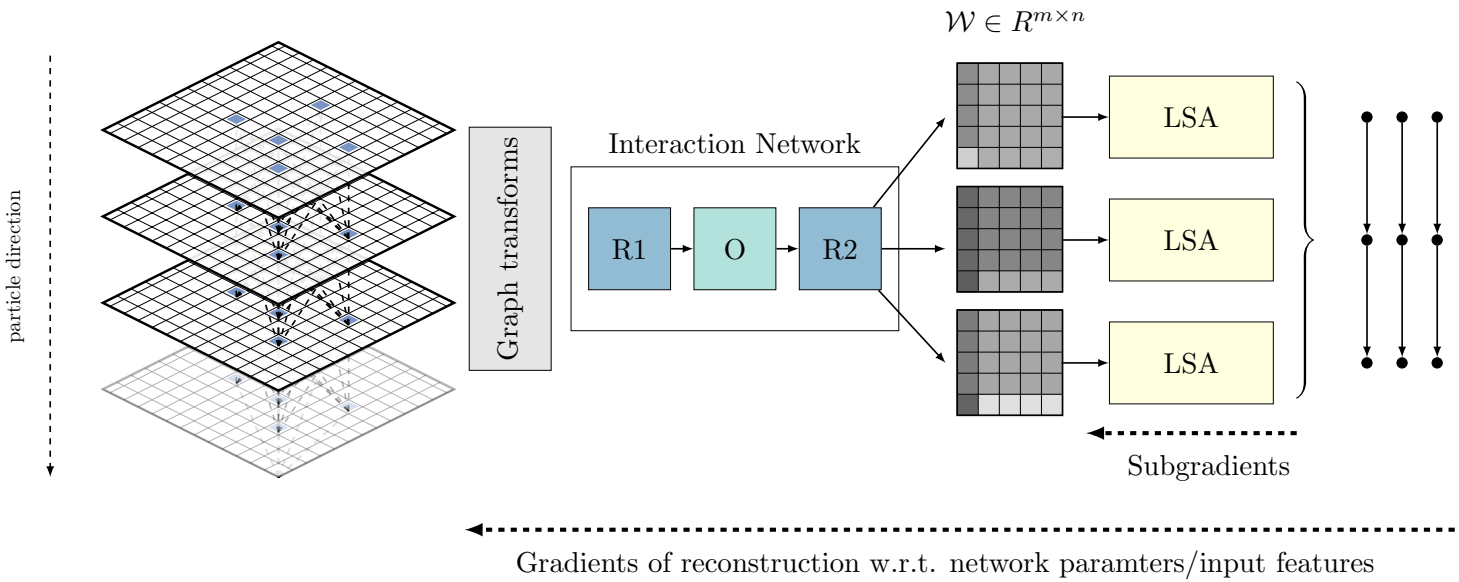


Image courtesy: Max Aehle et al. Exploration of differentiability in a proton computed tomography simulation framework. Phys Med Biol. 2023 Dec 15;68(24).

- **Question**: Is it possible to provide an E2E differentiable tracking algorithm leveraging SOTA edge classification GNNs? And how does it compare to its two-step counterpart?

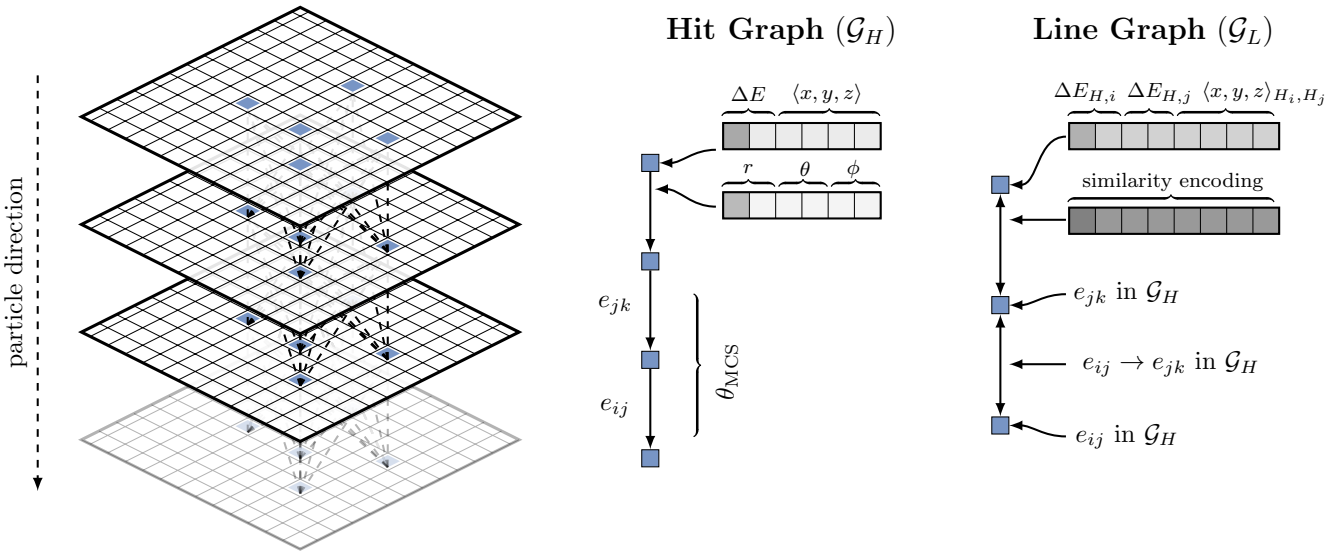
Graph Neural Network Architecture



- Prediction of edge scores on input graph, containing all particle hits
- Unique assignment of track segments using linear sum assignment (bipartite matching)
- End-to-end optimization using decision-focused learning $\rightarrow$ linearization of the solver mapping at the point $y(\hat{\mathcal{C}})$ using original and perturbed costs[2]

[2] Vlastelica, M., Paulus, A., Musil, V., Martius, G., Rolinek, M. (2020). Differentiation of Blackbox Combinatorial Solvers. 8th International Conference on Learning Representations, ICLR 2020, 1–19.

Graph Construction



- Transform edges in hit graph to nodes in line graph $\rightarrow$ scattering is parameterized by edges $\rightarrow$ strong inductive bias.
- Requires effective filtering of edges (threshold based on train set)

Blackbox Gradient Estimation for Combinatorial Solver

- Vlastelica et al. [3] define a general blackbox differentiation scheme for combinatorial solvers of form $y(w) = \arg\min_{y \in \mathcal{Y}} c(w, y)$ by considering the linearization of the solver mapping at the point $y(\hat{w})$ according to

$$\nabla_w f_\lambda(\hat{w}) := -\frac{1}{\lambda}\left[y(\hat{w}) - y_\lambda(w')\right], \quad \text{where} \quad w' = \text{clip}\left(\hat{w} + \lambda \frac{dL}{dy}(y(\hat{w})), 0, \infty\right) \tag{1}$$

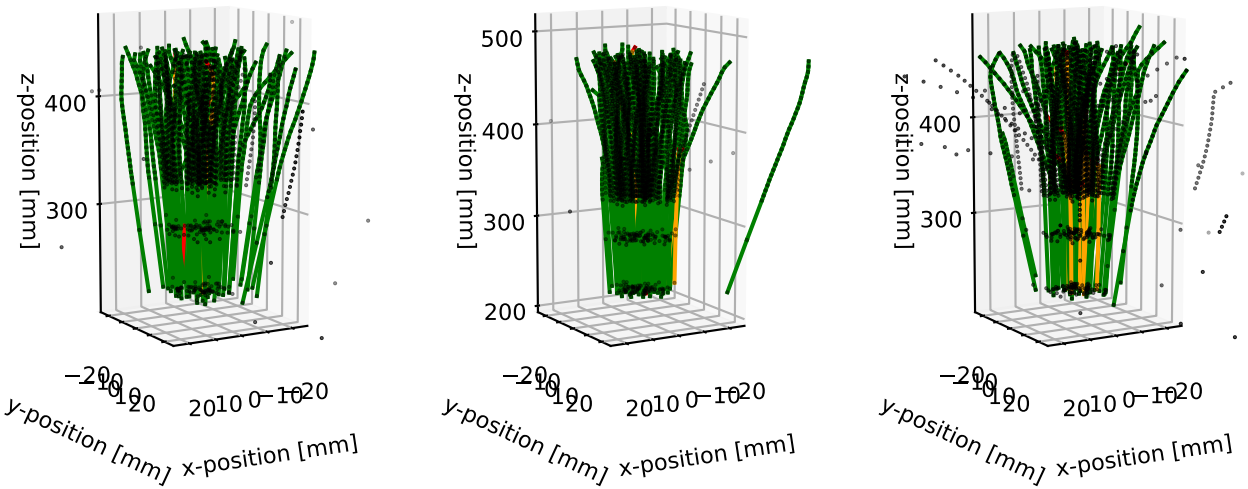
- **Advantages**:
  - We can remain with exact solvers without the necessity of using any relaxation of the combinatorial problem.
  - Simple implementation (requires only a second evaluation of the solver in the backward pass).

[3] Vlastelica, M., Paulus, A., Musil, V., Martius, G., & Rolinek, M. (2020). Differentiation of Blackbox Combinatorial Solvers. 8th International Conference on Learning Representations, ICLR 2020, 1–19.

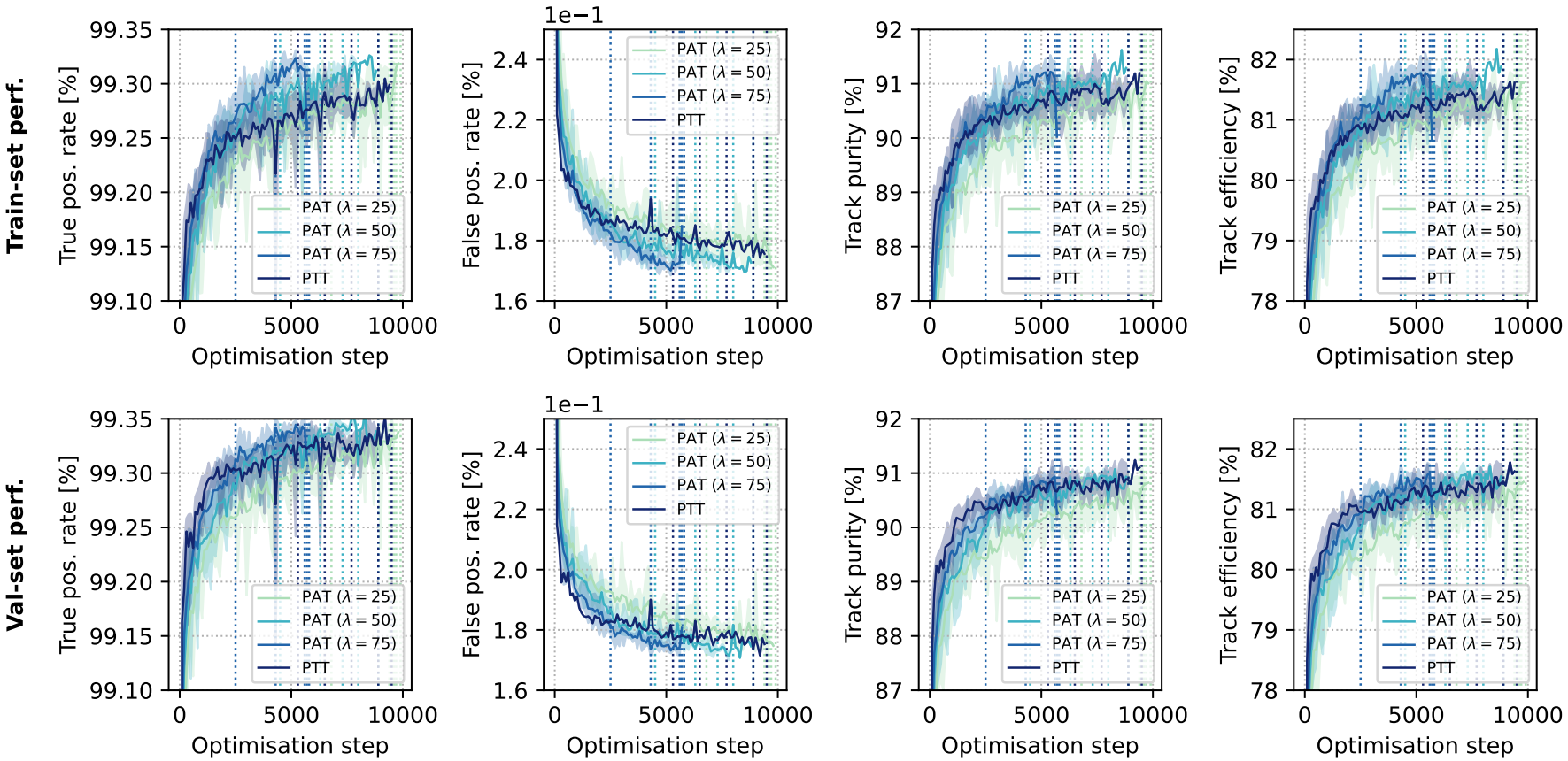
Network Optimization and Evaluation

1. **Predict-then-track (PTT):**
   - Minimize BCE-loss of edge scores.
2. **Predict-and-track (PTT):**
   - Minimize Hamming loss of edge assignments.

- MC simulated data for 100, 150 and 200 mm water phantom and 230 MeV pencil beam:
  - Train: 100k particles ($100 p^+/F$)
  - Eval: 5k particles ($100 p^+/F$)
  - Test: 10k particles (50-150 p^+/F)



Training Performance



Reconstruction Performance

		100 mm Water Phantom		150 mm Water Phantom		200 mm Water Phantom	
p^+/F	Algorithm	p [%] (↑)	ϵ [%] (↑)	p [%] (↑)	ϵ [%] (↑)	p [%] (↑)	ϵ [%] (↑)
50	Track follower[2]	87.1±0.0	78.1±0.0	89.4±0.0	81.1±0.0	90.8±0.0	82.1±0.0
	GNN_{PTT} (BCE. edge score)	94.9±0.1	83.8±0.0	96.2±0.2	86.9±0.2	96.4±0.0	88.7±0.1
	GNN_{PAT} ($\lambda = 25.0$)	94.9±0.1	83.8±0.1	96.2±0.1	87.0±0.1	96.5±0.1	88.9±0.1
	GNN_{PAT} ($\lambda = 50.0$)	95.0±0.2	83.9±0.2	96.3±0.2	87.1±0.2	96.5±0.1	89.0±0.1
	GNN_{PAT} ($\lambda = 75.0$)	95.0±0.2	83.9±0.2	96.3±0.1	87.1±0.1	96.5±0.0	89.0±0.0
100	Track follower[2]	80.6±0.0	71.7±0.0	84.7±0.0	76.4±0.0	85.8±0.0	77.5±0.0
	GNN_{PTT} (BCE. edge score)	87.4±0.3	75.2±0.2	91.9±0.3	82.4±0.3	91.7±0.2	83.8±0.1
	GNN_{PAT} ($\lambda = 25.0$)	87.3±0.2	75.0±0.2	91.7±0.2	82.1±0.3	92.2±0.2	84.4±0.2
	GNN_{PAT} ($\lambda = 50.0$)	87.4±0.3	75.1±0.2	92.0±0.1	82.4±0.2	92.5±0.1	84.6±0.1
	GNN_{PAT} ($\lambda = 75.0$)	87.4±0.2	75.1±0.2	91.9±0.1	82.4±0.1	92.3±0.2	84.4±0.2
150	Track follower [2]	75.6±0.0	67.2±0.0	80.1±0.1	72.2±0.0	82.5±0.0	74.6±0.0
	GNN_{PTT} (BCE. edge score)	77.5±0.4	65.0±0.3	84.9±0.3	75.3±0.3	87.2±0.2	79.6±0.2
	GNN_{PAT} ($\lambda = 25.0$)	76.7±0.2	64.3±0.2	84.8±0.3	75.1±0.3	87.6±0.3	80.1±0.3
	GNN_{PAT} ($\lambda = 50.0$)	76.8±0.3	64.4±0.3	85.1±0.2	75.5±0.2	88.1±0.4	80.6±0.4
	GNN_{PAT} ($\lambda = 75.0$)	76.6±0.1	64.3±0.1	85.0±0.2	75.4±0.2	87.8±0.2	80.4±0.2

[2] Pettersen, H., Meric, I., Odland, O., Shafiee, H., Sølie, J., Röhrich, D. (2020). Proton tracking algorithm in a pixel-based range telescope for proton computed tomography.

E2E Differentiable Particle Tracking from a Loss Landscape Perspective

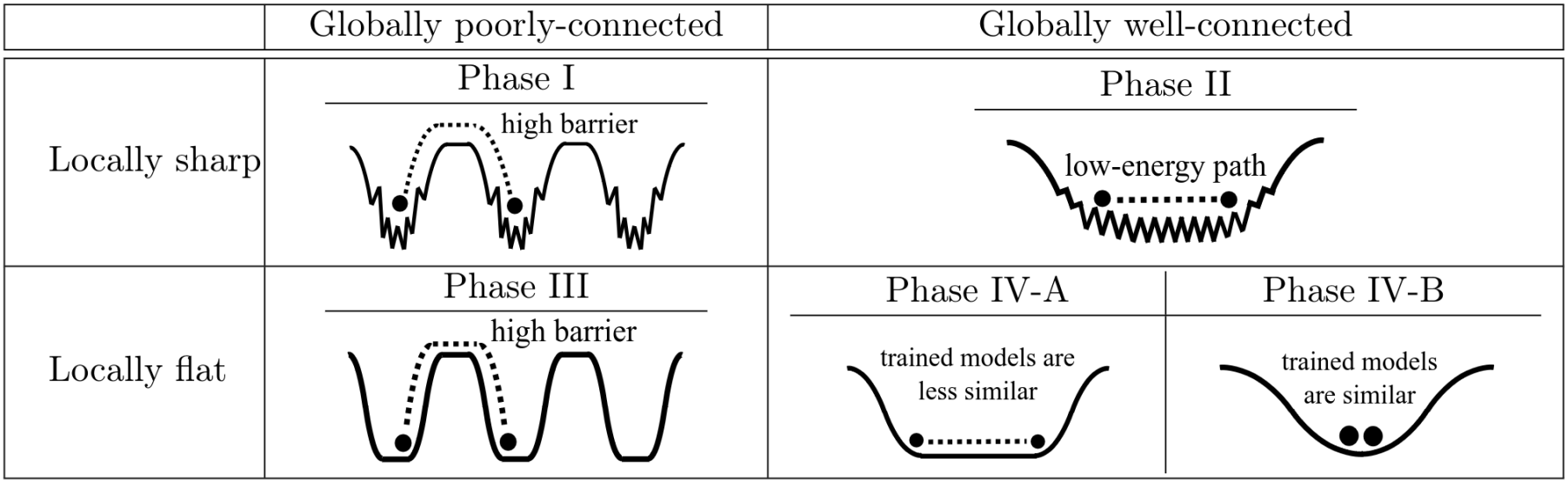


Figure taken from Yang et al., 2021[4]

[4] Yaoqing Yang, Liam Hodgkinson, Ryan Theisen, Joe Zou, Joseph E. Gonzalez, Kannan Ramchandran, Michael W. Mahoney (2021). Taxonomizing local versus global structure in neural network loss landscapes. In Advances in Neural Information

E2E Differentiable Particle Tracking from a Loss Landscape Perspective

Loss Surfaces

- Visualization of loss in landscape at minima using a 2D projection along vectors ν and η:

$$f(\alpha, \beta) = L(\theta^* + \alpha\nu + \beta\eta), \tag{2}$$

- We chose the eigenvectors of the hessian matrix with the highest eigenvalue. Projection along the highest curvature of the loss landscape.
- Characteristic behavior of loss landscape is consistent across different directions[a].

[a] Li, H., Xu, Z., Taylor, G., Studer, C., Goldstein, T. (2018). Visualizing the loss landscape of neural nets. Advances in Neural Information Processing Systems, 6389–6399.

Representational Similarities (CKA-Similarity)

- Correlation-like similarity metric [a] to compare learned representations of neural network layers of same or different models.
- Uses the Hilbert-Schmidt Independence Criterion (HSIC)

$$\mathrm{CKA}(K, L) = \frac{\mathrm{HSIC}(K, L)}{\sqrt{\mathrm{HSIC}(L, L)\mathrm{HSIC}(K, K)}}. \tag{3}$$

[a] Simon Kornblith, Mohammad Norouzi, Honglak Lee, Geoffrey Hinton. (2019). Similarity of Neural Network Representations Revisited.

E2E Differentiable Particle Tracking from a Loss Landscape Perspective

Functional Similarities

- Prediction instability or churn captures the average ratio of disagreement between predictions of different models f_1 and f_2[a]:

$$d(f_1, f2) = \mathbb{E}_{x,f_1,f_2} \mathbb{1} \left\{ \arg\max f_1(x) \neq \arg\max f_2(x) \right\}. \tag{4}$$

- We calculate the min-max normalized to provide a better intuition.

[a] Mahdi Milani Fard, Quentin Cormier, Kevin Canini, Maya Gupta (2016). Launch and Iterate: Reducing Prediction Churn. Advances in Neural Information Processing Systems, 29

Mode Connectivity

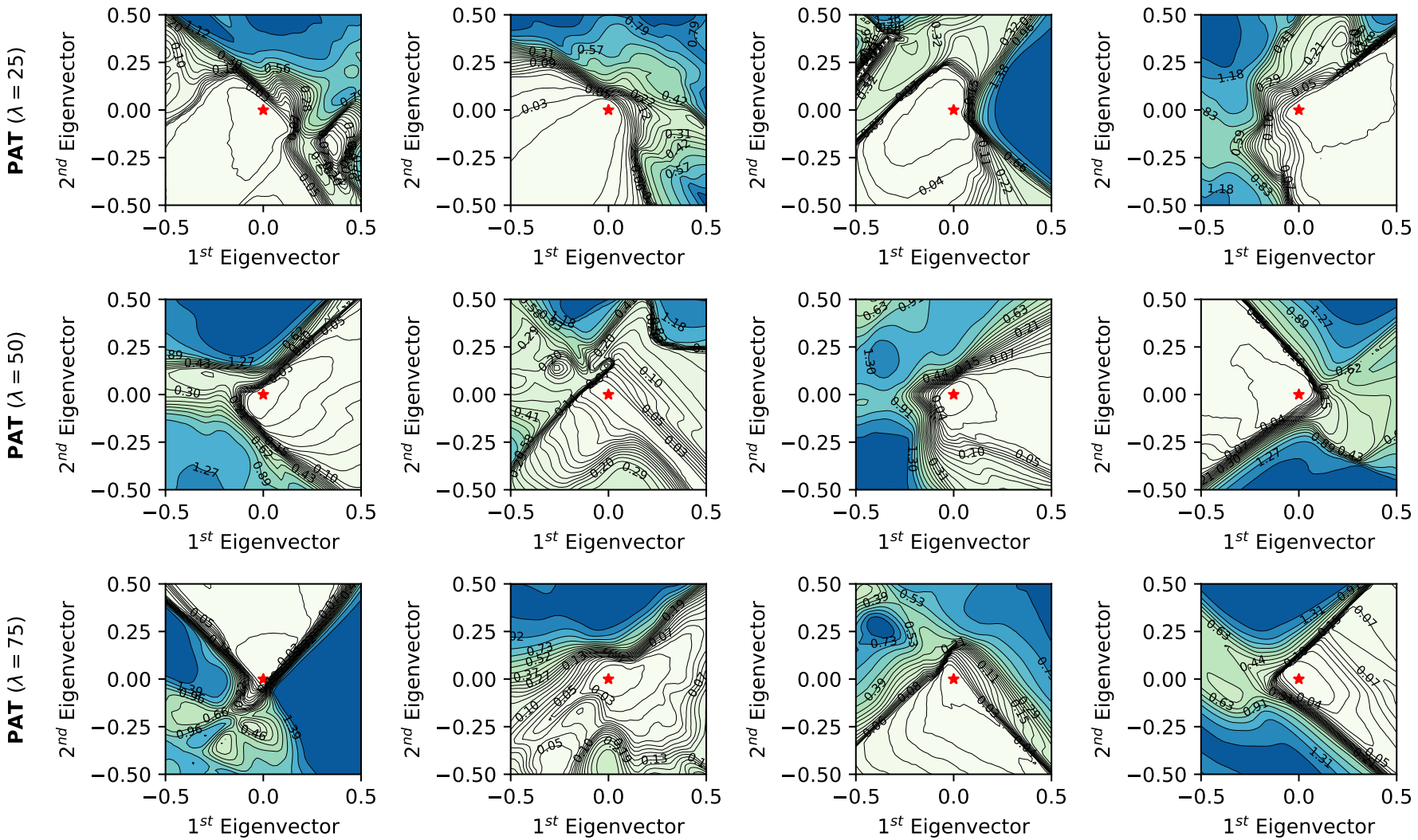
- Given two mode, we aim to find a low-energy connecting curve $\psi_\theta(t) : [0, 1] \to \mathbb{R}^d$, defined as the integral loss alongside the parameterized curve[a]:

$$\mathcal{L}(\theta) = \int_0^1 \mathcal{L}(\psi_\theta(t))dt = \mathbb{E}_{t\sim U(0,1)} \left[\mathcal{L}\left(\psi_\theta(t)\right) \right]. \tag{5}$$

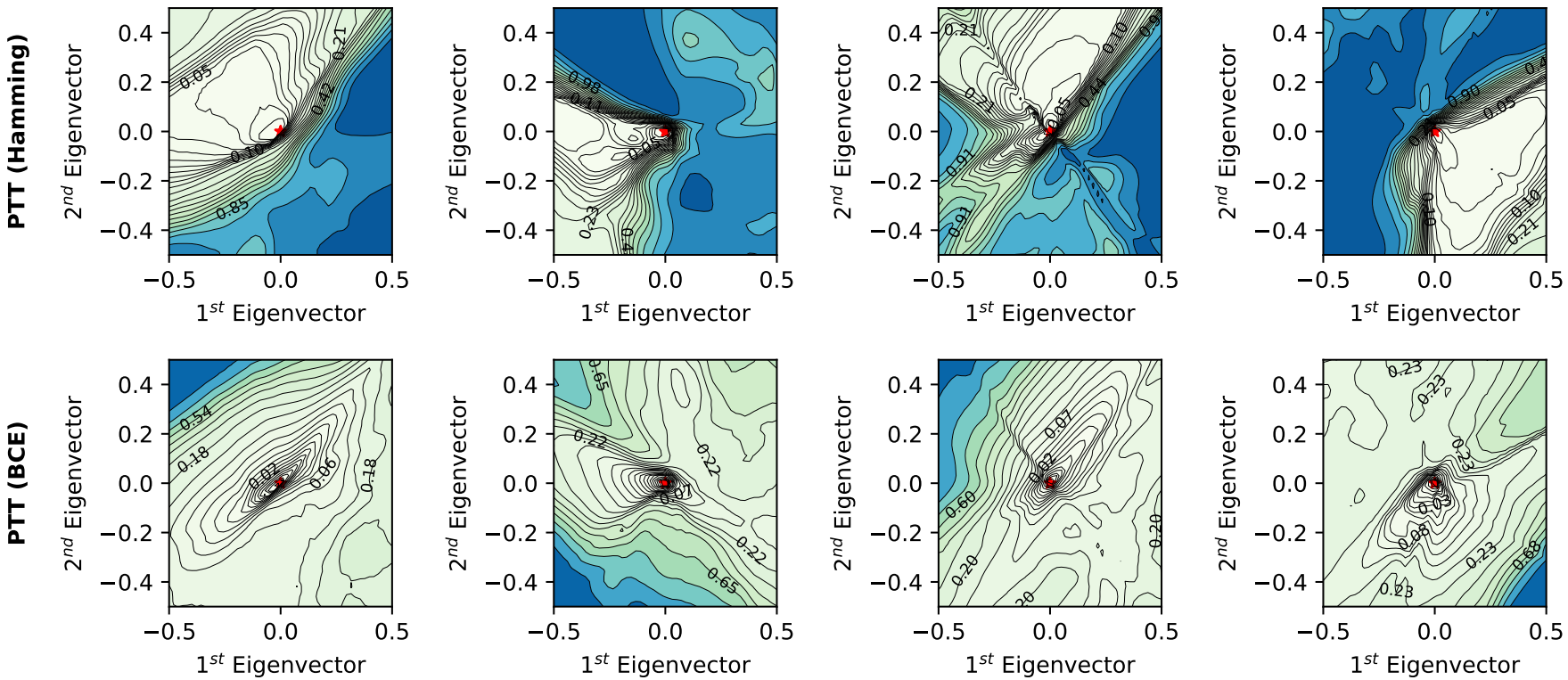
1. Non-linear connectivity (Bézier curves) $\to$ intuition about the general connectivity.
2. Linear connectivity $\to$ Intuition about the mechanistic similarity.

[a] Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D., Wilson, A. (2018). Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs. In Advances in Neural Information Processing Systems. Curran Associates, Inc..

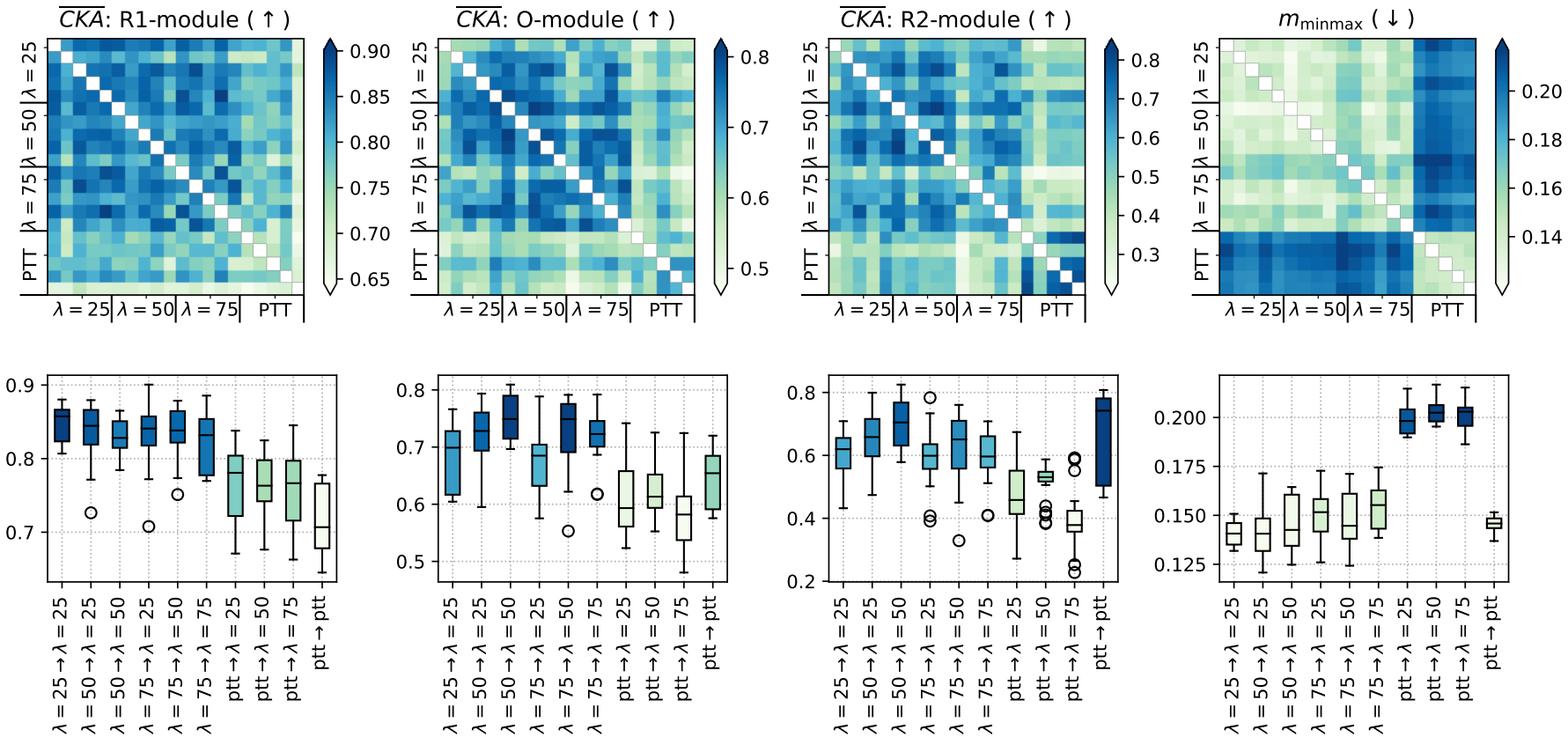
Local Structure of the Task Loss Surfaces for Decision-focused Learning



Agreement of Prediction Loss and Task Loss Surfaces



Functional and Representational Similarities



Mode Connectivity

		Hamming loss [$\times 1e^{-3}$] (↓)			True positive rate [%] (↑)		
Model	Curve	Min	Max	Mean	Min	Max	Mean
PAT ($\lambda = 25$)	Linear	6.64±0.09	228.58±110.62	37.73±19.92	75.01±11.77	99.32±0.01	95.80±2.16
PAT ($\lambda = 25$)	Bézier (Ham.)	6.64±0.07	8.14±0.53	7.11±0.10	99.12±0.09	99.32±0.01	99.26±0.01
PAT ($\lambda = 25$)	Bézier (BCE)	6.65±0.11	7.78±0.13	7.35±0.06	99.18±0.02	99.32±0.01	99.24±0.01
PAT ($\lambda = 50$)	Linear	6.47±0.12	251.09±102.46	43.46±22.46	71.57±10.24	99.34±0.01	95.01±2.39
PAT ($\lambda = 50$)	Bézier (Ham.)	6.46±0.12	8.01±0.40	7.00±0.09	99.15±0.05	99.34±0.01	99.28±0.01
PAT ($\lambda = 50$)	Bézier (BCE)	6.47±0.08	7.76±0.12	7.35±0.05	99.18±0.02	99.34±0.01	99.24±0.01
PAT ($\lambda = 75$)	Linear	6.49±0.04	345.56±133.60	82.49±44.28	63.04±13.58	99.34±0.00	91.13±4.60
PAT ($\lambda = 75$)	Bézier (Ham.)	6.47±0.05	8.22±0.79	7.02±0.16	99.11±0.12	99.34±0.00	99.27±0.02
PAT ($\lambda = 75$)	Bézier (BCE)	6.51±0.04	8.12±0.72	7.40±0.08	99.16±0.05	99.34±0.00	99.23±0.01
PTT	Linear	6.70±0.07	332.39±116.90	107.24±42.20	63.67±11.94	99.32±0.01	87.59±4.31
PTT	Bézier (Ham.)	6.70±0.07	15.33±5.54	9.56±1.07	97.98±0.94	99.32±0.01	98.91±0.18
PTT	Bézier (BCE)	6.69±0.07	7.60±0.21	7.07±0.06	99.21±0.03	99.32±0.01	99.28±0.01

		Hamming loss [$\times 1e^{-3}$] (↓)			True positive rate [%] (↑)		
Model	Curve	Min	Max	Mean	Min	Max	Mean
$GNN_{PTT} \rightarrow GNN_{PAT}$	Linear	6.60±0.13	337.23±169.20	117.42±76.26	63.90±17.66	99.33±0.01	87.12±8.06
$GNN_{PTT} \rightarrow GNN_{PAT}$	Bézier (BCE)	6.60±0.13	11.32±3.25	8.06±0.58	98.65±0.52	99.33±0.01	99.13±0.08
$GNN_{PTT} \rightarrow GNN_{PAT}$	Bézier (Ham.)	6.59±0.12	7.75±0.36	7.19±0.09	99.19±0.03	99.33±0.01	99.26±0.01

Mode Connectivity - Global Connectivity

		Hamming loss [$\times 1e^{-3}$] (↓)			True positive rate [%] (↑)		
Model	Curve	Min	Max	Mean	Min	Max	Mean
PAT ($\lambda = 25$)	Linear	6.64±0.09	228.58±110.62	37.73±19.92	75.01±11.77	99.32±0.01	95.80±2.16
PAT ($\lambda = 25$)	Bézier (Ham.)	6.64±0.07	8.14±0.53	7.11±0.10	99.12±0.09	99.32±0.01	99.26±0.01
PAT ($\lambda = 25$)	Bézier (BCE)	6.65±0.11	7.78±0.13	7.35±0.06	99.18±0.02	99.32±0.01	99.24±0.01
PAT ($\lambda = 50$)	Linear	6.47±0.12	251.09±102.46	43.46±22.46	71.57±10.24	99.34±0.01	95.01±2.39
PAT ($\lambda = 50$)	Bézier (Ham.)	6.46±0.12	8.01±0.40	7.00±0.09	99.15±0.05	99.34±0.01	99.28±0.01
PAT ($\lambda = 50$)	Bézier (BCE)	6.47±0.08	7.76±0.12	7.35±0.05	99.18±0.02	99.34±0.01	99.24±0.01
PAT ($\lambda = 75$)	Linear	6.49±0.04	345.56±133.60	82.49±44.28	63.04±13.58	99.34±0.00	91.13±4.60
PAT ($\lambda = 75$)	Bézier (Ham.)	6.47±0.05	8.22±0.79	7.02±0.16	99.11±0.12	99.34±0.00	99.27±0.02
PAT ($\lambda = 75$)	Bézier (BCE)	6.51±0.04	8.12±0.72	7.40±0.08	99.16±0.05	99.34±0.00	99.23±0.01
PTT	Linear	6.70±0.07	332.39±116.90	107.24±42.20	63.67±11.94	99.32±0.01	87.59±4.31
PTT	Bézier (Ham.)	6.70±0.07	15.33±5.54	9.56±1.07	97.98±0.94	99.32±0.01	98.91±0.18
PTT	Bézier (BCE)	6.69±0.07	7.60±0.21	7.07±0.06	99.21±0.03	99.32±0.01	99.28±0.01

		Hamming loss [$\times 1e^{-3}$] (↓)			True positive rate [%] (↑)		
Model	Curve	Min	Max	Mean	Min	Max	Mean
$GNN_{PTT} \rightarrow GNN_{PAT}$	Linear	6.60±0.13	337.23±169.20	117.42±76.26	63.90±17.66	99.33±0.01	87.12±8.06
$GNN_{PTT} \rightarrow GNN_{PAT}$	Bézier (Ham.)	6.60±0.13	11.32±3.25	8.06±0.58	98.65±0.52	99.33±0.01	99.13±0.08
$GNN_{PTT} \rightarrow GNN_{PAT}$	Bézier (BCE.)	6.59±0.12	7.75±0.36	7.19±0.09	99.19±0.03	99.33±0.01	99.26±0.01

Low-energy connecting curves for all configurations & for combination of PAT and PTT

Mode Connectivity - Mechanistic Similarity

		Hamming loss [$\times 1e^{-3}$] (↓)			True positive rate [%] (↑)		
Model	Curve	Min	Max	Mean	Min	Max	Mean
PAT ($\lambda = 25$)	Linear	6.64±0.09	228.58±110.62	37.73±19.92	75.01±11.77	99.32±0.01	95.80±2.16
PAT ($\lambda = 25$)	Bézier (Ham.)	6.64±0.07	8.14±0.53	7.11±0.10	99.12±0.09	99.32±0.01	99.26±0.01
PAT ($\lambda = 25$)	Bézier (BCE)	6.65±0.11	7.78±0.13	7.35±0.06	99.18±0.02	99.32±0.01	99.24±0.01
PAT ($\lambda = 50$)	Linear	6.47±0.12	251.09±102.46	43.46±22.46	71.57±10.24	99.34±0.01	95.01±2.39
PAT ($\lambda = 50$)	Bézier (Ham.)	6.46±0.12	8.01±0.40	7.00±0.09	99.15±0.05	99.34±0.01	99.28±0.01
PAT ($\lambda = 50$)	Bézier (BCE)	6.47±0.08	7.76±0.12	7.35±0.05	99.18±0.02	99.34±0.01	99.24±0.01
PAT ($\lambda = 75$)	Linear	6.49±0.04	345.56±133.60	82.49±44.28	63.04±13.58	99.34±0.00	91.13±4.60
PAT ($\lambda = 75$)	Bézier (Ham.)	6.47±0.05	8.22±0.79	7.02±0.16	99.11±0.12	99.34±0.00	99.27±0.02
PAT ($\lambda = 75$)	Bézier (BCE)	6.51±0.04	8.12±0.72	7.40±0.08	99.16±0.05	99.34±0.00	99.23±0.01
PTT	Linear	6.70±0.07	332.39±116.90	107.24±42.20	63.67±11.94	99.32±0.01	87.59±4.31
PTT	Bézier (Ham.)	6.70±0.07	15.33±5.54	9.56±1.07	97.98±0.94	99.32±0.01	98.91±0.18
PTT	Bézier (BCE)	6.69±0.07	7.60±0.21	7.07±0.06	99.21±0.03	99.32±0.01	99.28±0.01

		Hamming loss [$\times 1e^{-3}$] (↓)			True positive rate [%] (↑)		
Model	Curve	Min	Max	Mean	Min	Max	Mean
$GNN_{PTT} \rightarrow GNN_{PAT}$	Linear	6.60±0.13	337.23±169.20	117.42±76.26	63.90±17.66	99.33±0.01	87.12±8.06
$GNN_{PTT} \rightarrow GNN_{PAT}$	Bézier (BCE)	6.60±0.13	11.32±3.25	8.06±0.58	98.65±0.52	99.33±0.01	99.13±0.08
$GNN_{PTT} \rightarrow GNN_{PAT}$	Bézier (Ham.)	6.59±0.12	7.75±0.36	7.19±0.09	99.19±0.03	99.33±0.01	99.26±0.01

Lack of linear connectivity → low mechanistic similarity

Mode Connectivity - Impact of Interpolation Factor λ

		Hamming loss [$\times 1e^{-3}$] (↓)			True positive rate [%] (↑)		
Model	Curve	Min	Max	Mean	Min	Max	Mean
PAT ($\lambda = 25$)	Linear	6.64±0.09	228.58±110.62	37.73±19.92	75.01±11.77	99.32±0.01	95.80±2.16
PAT ($\lambda = 25$)	Bézier (Ham.)	6.64±0.07	8.14±0.53	7.11±0.10	99.12±0.09	99.32±0.01	99.26±0.01
PAT ($\lambda = 25$)	Bézier (BCE)	6.65±0.11	7.78±0.13	7.35±0.06	99.18±0.02	99.32±0.01	99.24±0.01
PAT ($\lambda = 50$)	Linear	6.47±0.12	251.09±102.46	43.46±22.46	71.57±10.24	99.34±0.01	95.01±2.39
PAT ($\lambda = 50$)	Bézier (Ham.)	6.46±0.12	8.01±0.40	7.00±0.09	99.15±0.05	99.34±0.01	99.28±0.01
PAT ($\lambda = 50$)	Bézier (BCE)	6.47±0.08	7.76±0.12	7.35±0.05	99.18±0.02	99.34±0.01	99.24±0.01
PAT ($\lambda = 75$)	Linear	6.49±0.04	345.56±133.60	82.49±44.28	63.04±13.58	99.34±0.00	91.13±4.60
PAT ($\lambda = 75$)	Bézier (Ham.)	6.47±0.05	8.22±0.79	7.02±0.16	99.11±0.12	99.34±0.00	99.27±0.02
PAT ($\lambda = 75$)	Bézier (BCE)	6.51±0.04	8.12±0.72	7.40±0.08	99.16±0.05	99.34±0.00	99.23±0.01
PTT	Linear	6.70±0.07	332.39±116.90	107.24±42.20	63.67±11.94	99.32±0.01	87.59±4.31
PTT	Bézier (Ham.)	6.70±0.07	15.33±5.54	9.56±1.07	97.98±0.94	99.32±0.01	98.91±0.18
PTT	Bézier (BCE)	6.69±0.07	7.60±0.21	7.07±0.06	99.21±0.03	99.32±0.01	99.28±0.01

		Hamming loss [$\times 1e^{-3}$] (↓)			True positive rate [%] (↑)		
Model	Curve	Min	Max	Mean	Min	Max	Mean
$GNN_{PTT} \rightarrow GNN_{PAT}$	Linear	6.60±0.13	337.23±169.20	117.42±76.26	63.90±17.66	99.33±0.01	87.12±8.06
$GNN_{PTT} \rightarrow GNN_{PAT}$	Bézier (BCE)	6.60±0.13	11.32±3.25	8.06±0.58	98.65±0.52	99.33±0.01	99.13±0.08
$GNN_{PTT} \rightarrow GNN_{PAT}$	Bézier (Ham.)	6.59±0.12	7.75±0.36	7.19±0.09	99.19±0.03	99.33±0.01	99.26±0.01

Increasing linear connectivity with decreasing interpolation factor $\lambda \rightarrow$ selection of sufficiently small λ values

Conclusion and Outlook

- E2E particle tracking is straightforward and works efficiently, providing results on-par with two-step approach while requiring less training iterations (better for smaller λ-values).
- Optimization works similar to two-step approaches, converging into globally-well connected minima. **However**: Substantial prediction instabilities (better for smaller λ-values) $\rightarrow$ unpredictable effects on downstream tasks.
- Providing gradient information enables various extensions (at least theoretically):
  1. Integration of auxiliary losses of downstream tasks to generate robust models that perform well on downstream task.
  2. Integration into complex architectures and learning schemes (e.g., MARL)
  3. Optimization of detector geometries for design optimization.

The Bergen pCT Collaboration

- University of Bergen, Norway
- Helse Bergen, Norway
- Western Norway University of Applied Science, Bergen, Norway
- Wigner Research Center for Physics, Budapest, Hungary
- DKFZ, Heidelberg, Germany
- Saint Petersburg State University, Saint Petersburg, Russia
- Utrecht University, Netherlands
- RPE LTU, Kharkiv, Ukraine
- Suranaree University of Technology, Nakhon Ratchasima, Thailand
- China Three Gorges University, Yichang, China
- University of Applied Sciences Worms, Germany
- University of Oslo, Norway
- Eötvös Loránd University, Budapest, Hungary
- University of Kaiserslautern Landau, Germany

Contact: tobias.kortus@rptu.de