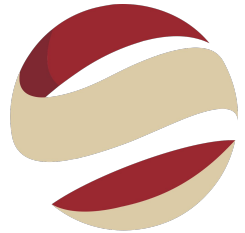
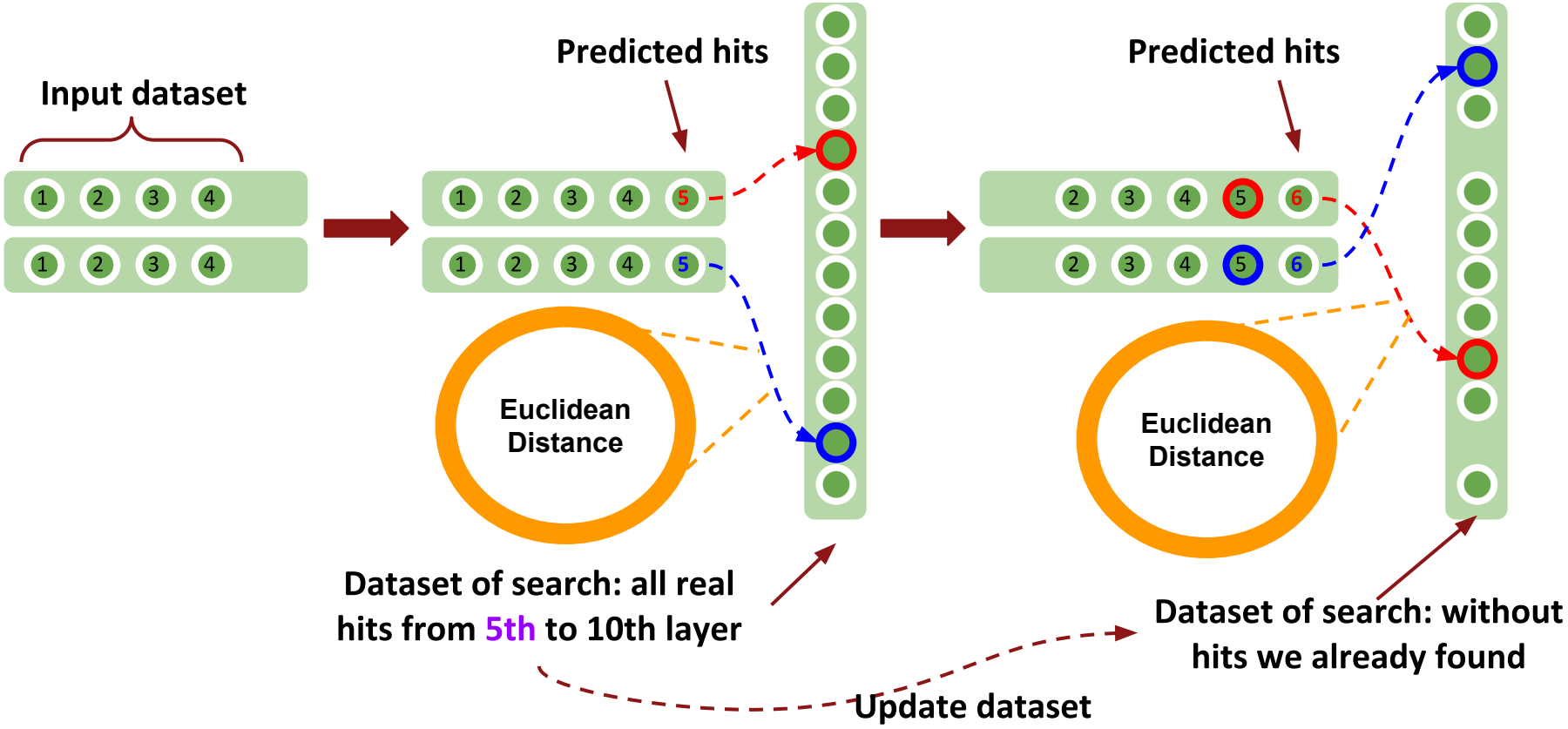


Track Reconstruction with BDT + Cosine Similarity

Angelo Santos - 24 / Sep / 2020

SPRACE

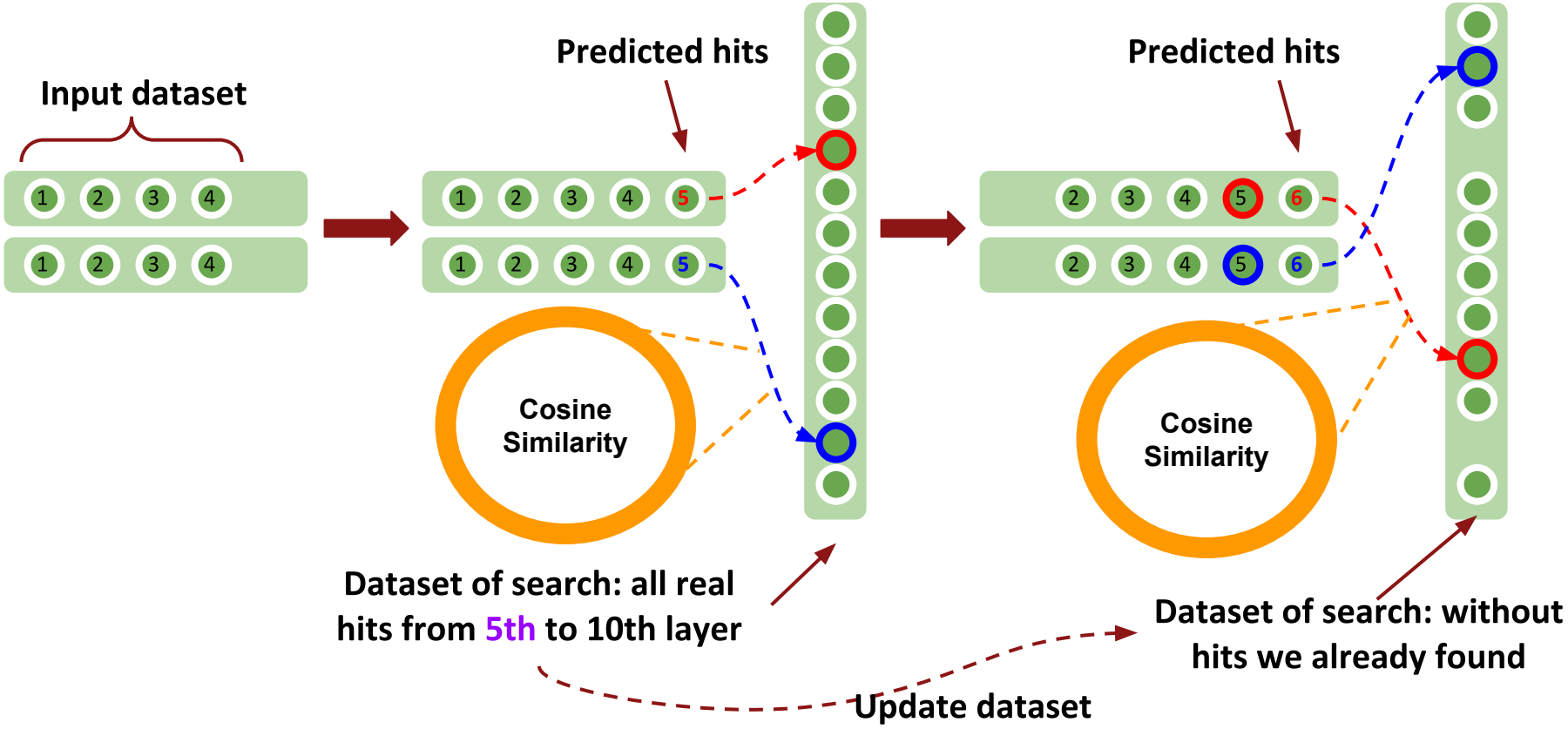
Application: Search via Euclidean Distance



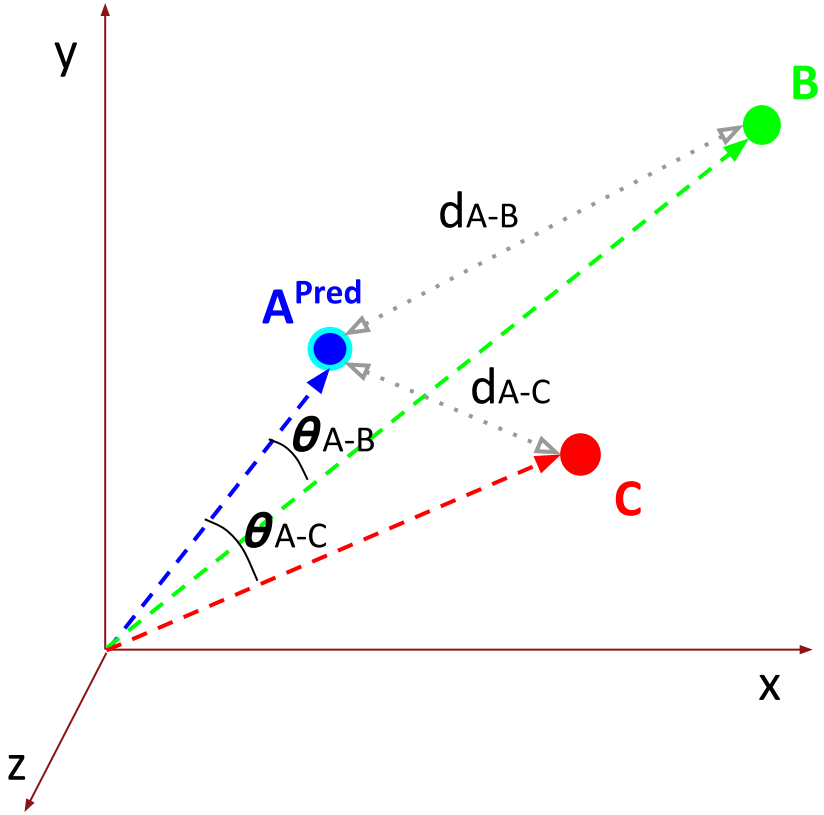
Previous Results: Successes and Failures

Euclidean Distance								
Update Dataset	**Layers**	**5**	**6**	**7**	**8**	**9**	**10**	**Whole Track**
Yes	**Success**	5026 (84%)	4815 (80%)	4626 (77%)	4272 (71%)	3793 (63%)	3552 (59%)	2411 (40%)
	Failure	974 (16%)	1185 (20%)	1374 (23%)	1728 (29%)	2207 (37%)	2448 (41%)	3589 (60%)
No	**Success**	5125 (85%)	5084 (85%)	5017 (84%)	4820 (80%)	4445 (74%)	4426 (74%)	3280 (55%)
	Failure	875 (15%)	916 (15%)	983 (16%)	1180 (20%)	1555 (26%)	1574 (26%)	2720 (45%)

Application: Search via Cosine Similarity



Cosine Similarity X Euclidean Distance



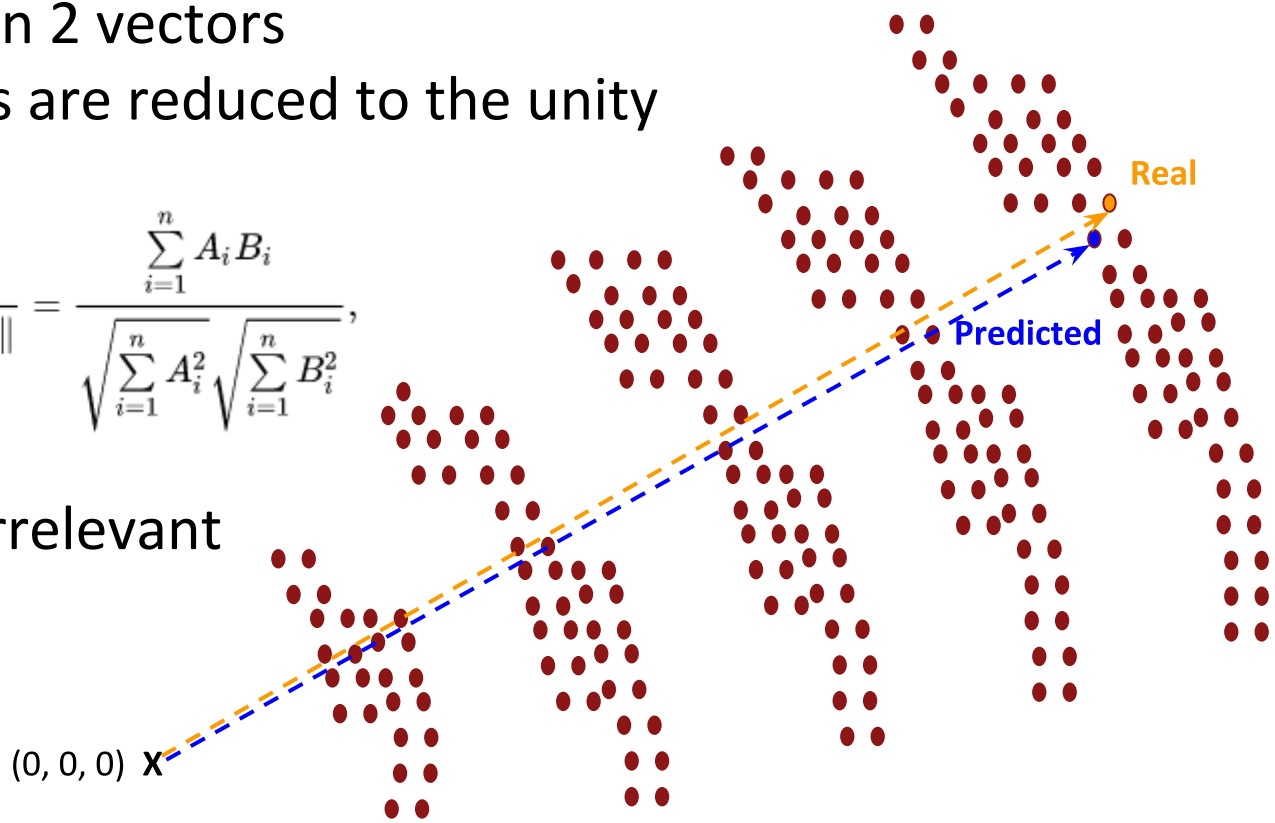
- With cosine similarity
  - $\cos(\theta_{A-B}) > \cos(\theta_{A-C})$
    - Then ⟶ choose B
- With Euclidean Distance
  - $d_{A-B} > d_{A-C}$
    - Then ⟶ choose C

Cosine Similarity

- Cosine between 2 vectors
- But the vectors are reduced to the unity

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

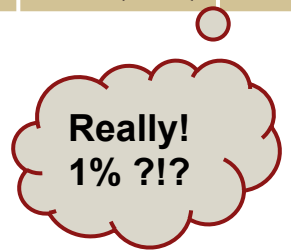
- Magnitude is irrelevant
- Angle is crucial



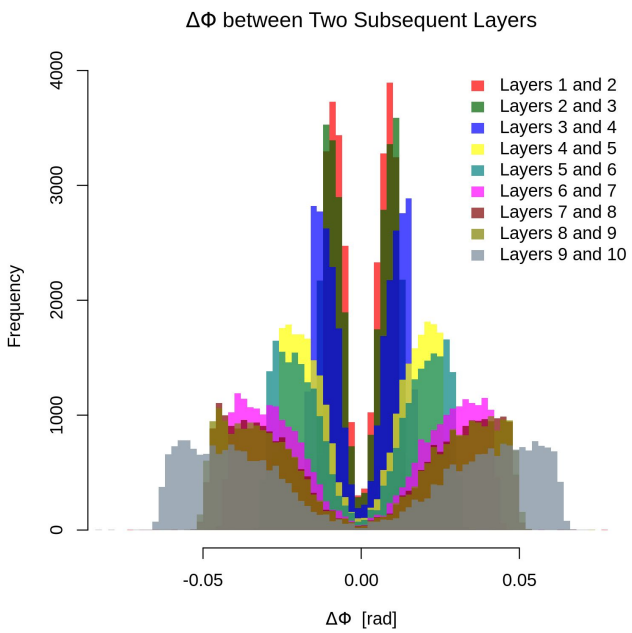
New Results: Successes and Failures

Euclidean Distance								
Update Dataset	**Layers**	**5**	**6**	**7**	**8**	**9**	**10**	**Whole Track**
Yes	**Success**	5026 (84%)	4815 (80%)	4626 (77%)	4272 (71%)	3793 (63%)	3552 (59%)	**2411 (40%)**
	Failure	974 (16%)	1185 (20%)	1374 (23%)	1728 (29%)	2207 (37%)	2448 (41%)	3589 (60%)

Cosine Similarity								
Update Dataset	**Layers**	**5**	**6**	**7**	**8**	**9**	**10**	**Whole Track**
Yes	**Success**	3461 (58%)	2502 (42%)	1741 (29%)	1037 (17%)	517 (9%)	117 (2%)	**76 (1%)**
	Failure	2539 (42%)	3498 (58%)	4259 (71%)	4963 (83%)	5483 (91%)	5883 (98%)	5924 (99%)



Displacing Vectors



Going to the outermost region, ΔΦ between 2 subsequent hits becomes larger

Real

Predicted

Track

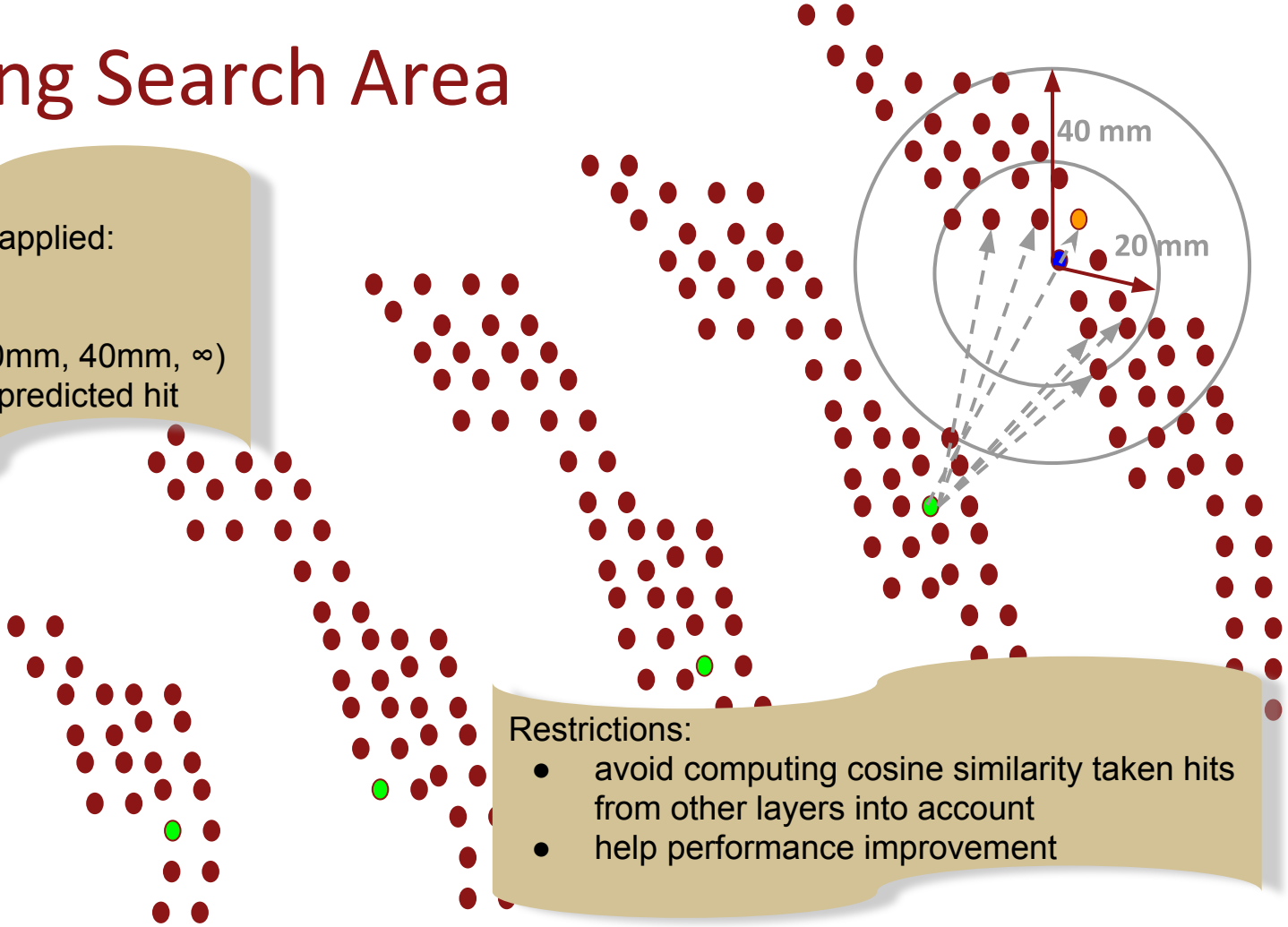
(0, 0, 0) X

All vectors must be displaced, so the new common origin is now the track hit in the fourth layer, rather than (0, 0, 0)

Restricting Search Area

3 restrictions are applied:

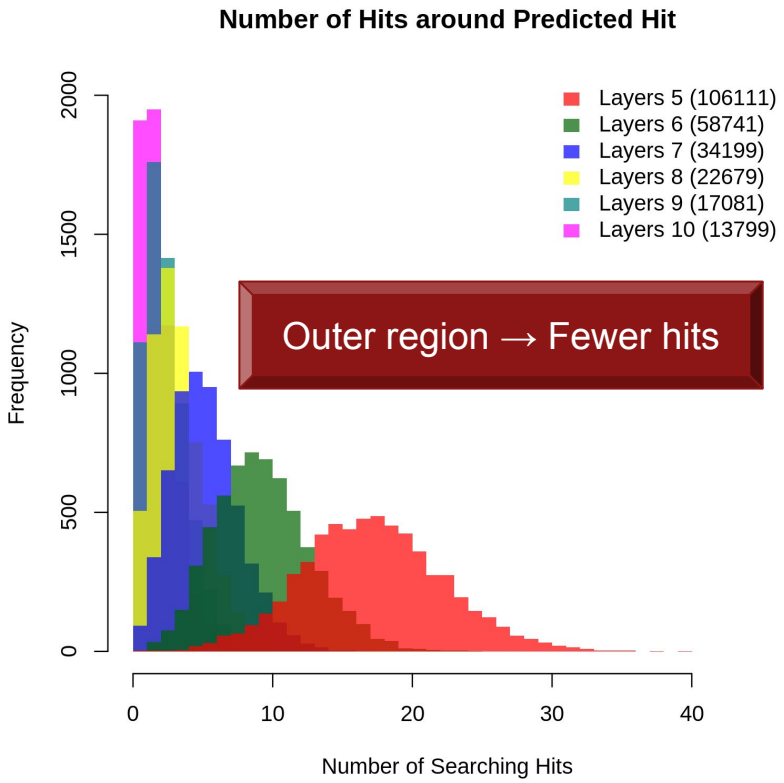
- volume ID
- layer ID
- radius < (20mm, 40mm, ∞) around the predicted hit

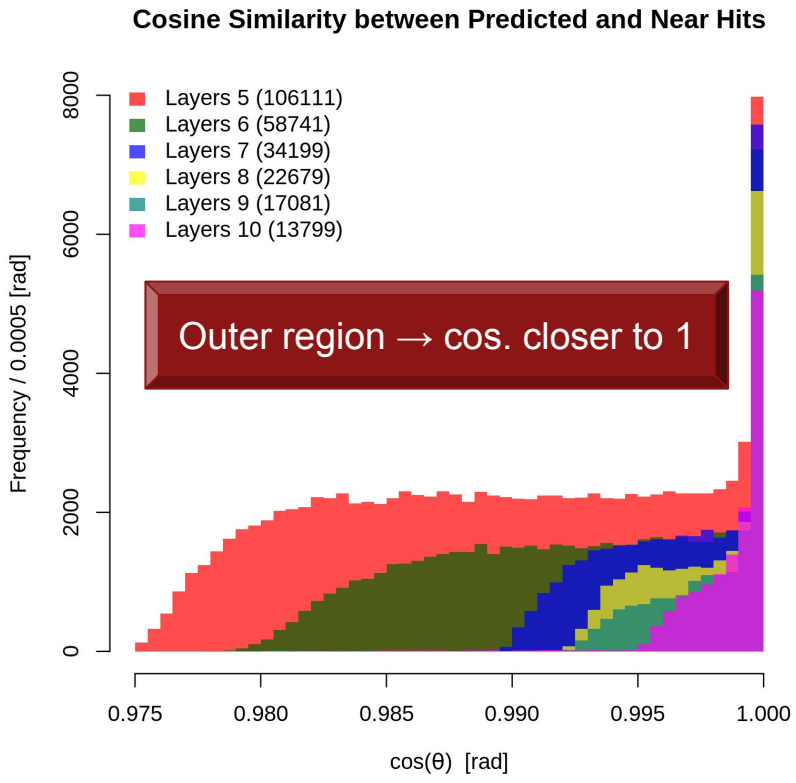


Restrictions:

- avoid computing cosine similarity taken hits from other layers into account
- help performance improvement

Cosine Similarity Distribution in Each Layer





Metrics: Successes and Failures

Euclidean Distance								
Update Dataset	**Layers**	**5**	**6**	**7**	**8**	**9**	**10**	**Whole Track**
Yes	**Success**	5026 (84%)	4815 (80%)	4626 (77%)	4272 (71%)	3793 (63%)	3552 (59%)	**2411 (40%)**
	Failure	974 (16%)	1185 (20%)	1374 (23%)	1728 (29%)	2207 (37%)	2448 (41%)	3589 (60%)
No	**Success**	5125 (85%)	5084 (85%)	5017 (84%)	4820 (80%)	4445 (74%)	4426 (74%)	**3280 (55%)**
	Failure	875 (15%)	916 (15%)	983 (16%)	1180 (20%)	1555 (26%)	1574 (26%)	2720 (45%)

Cosine Similarity								
Update Dataset	**Layers**	**5**	**6**	**7**	**8**	**9**	**10**	**Whole Track**
Yes	**Success**	5072 (85%)	4918 (82%)	4734 (79%)	4438 (74%)	4011 (67%)	3790 (63%)	**2684 (45%)**
	Failure	928 (15%)	1082 (18%)	1266 (21%)	1562 (26%)	1989 (33%)	2210 (37%)	3316 (55%)
No	**Success**	**5178 (86%)**	5202 (87%)	5132 (86%)	4956 (83%)	4570 (76%)	4495 (75%)	**3415 (57%)**
	Failure	822 (14%)	798 (13%)	868 (14%)	1044 (17%)	1430 (24%)	1505 (25%)	2585 (43%)

Conclusions

- Cosine similarity improves results from 1 to 5%
- Best result is 57% of success in the whole track reconstruction
- However, there is still room to improvement
  - May get better GBM model optimizing parameters 
    - with cosine similarity (instead of Euclidean distance)
  - May improve data searching with a (an “Inception”) 
    - Probability distribution
    - Machine Learning Technique
  - May find a way to deal with combinatorials 
    - Prediction of next hit based on 4, 5, 6, 7, 8, 9 hits
    - Other ML technique?

Links to Past Presentations

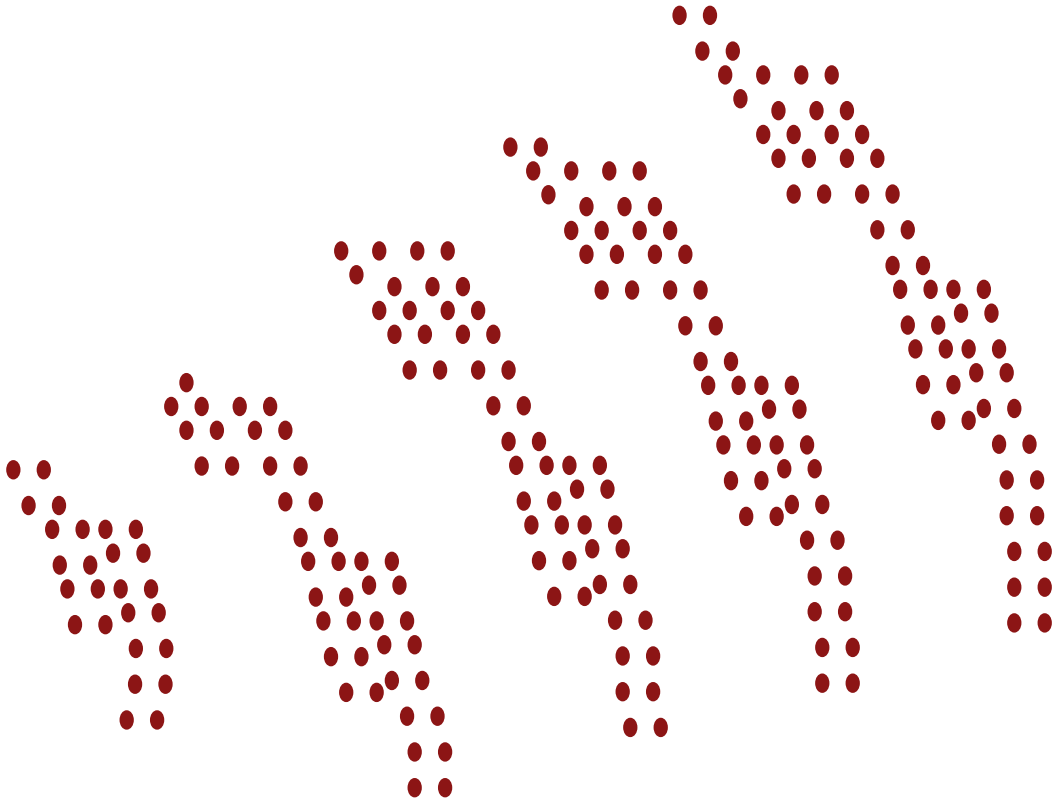
2019

- 24/Jul/2019
- 09/Aug/2019
- 23/Aug/2019
- 30/Aug/2019
- 27/Sep/2019

2020

- 22/Apr/2020
- 13/May/2020
- 20/May/2020
- 10/Jun/2020
- 24/Jun/2020
- 08/Jul/2020

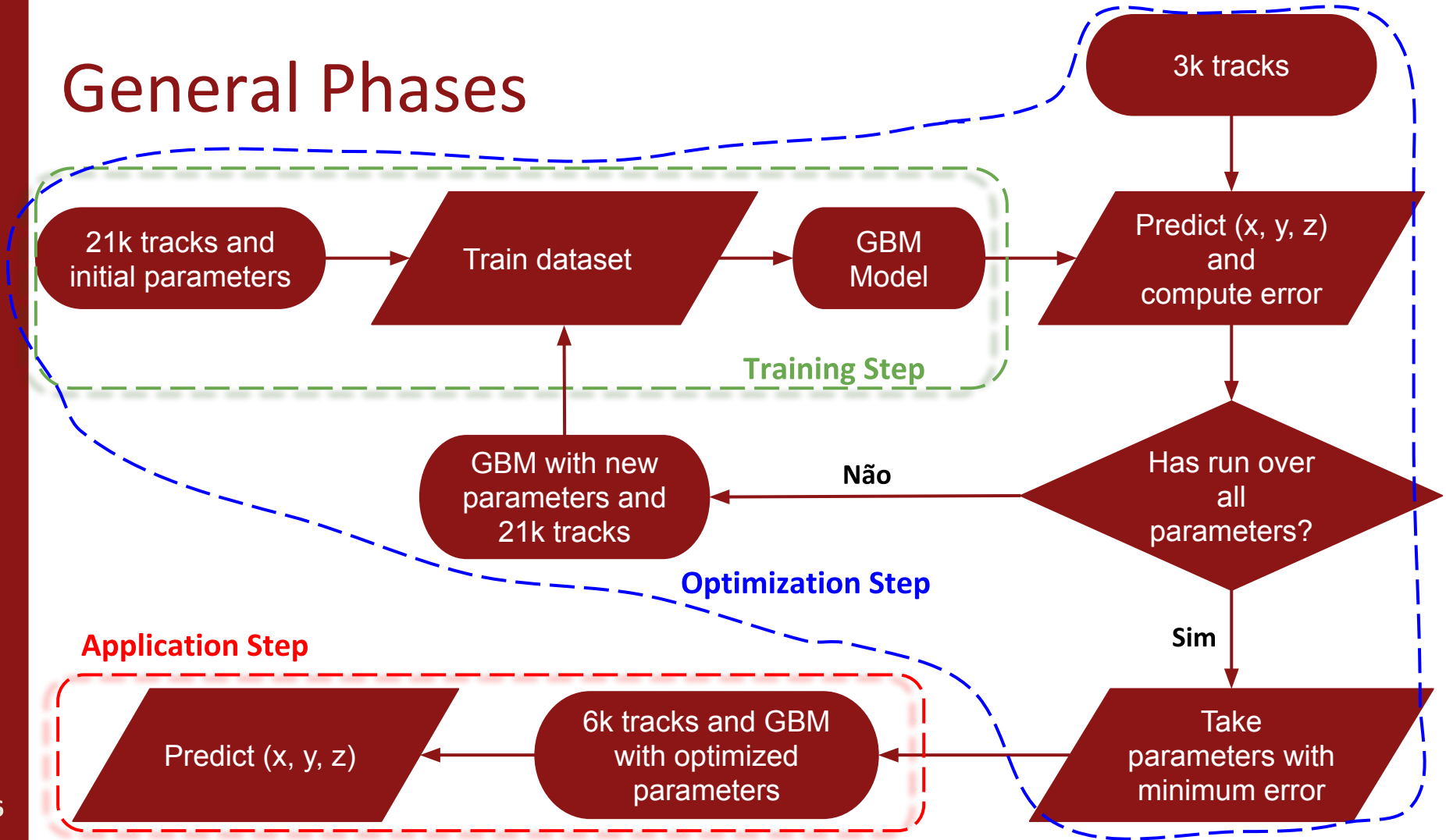
Backup



Data Set

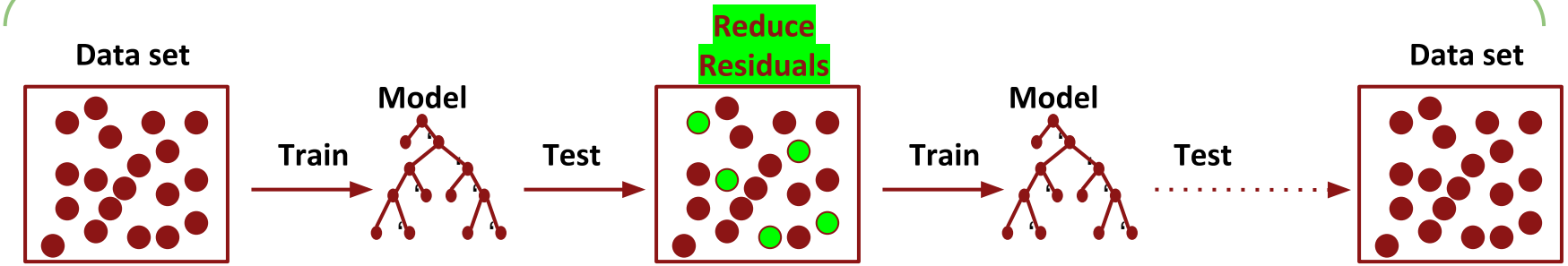
- /data/track-ml/eramia/results/eta_n0.5-0.5_phi_ninf-pinf.csv
  - 30k tracks
  - $p_T > 1$ GeV
  - $|\eta| < 0.5$
  - $|\phi| \rightarrow$ no cut
  - 10 hits/track
- Analysis
  - Training $\rightarrow$21k tracks (70%)
  - Optimization $\rightarrow$ 3k tracks (10%)
  - Application $\rightarrow$ 6k tracks

General Phases

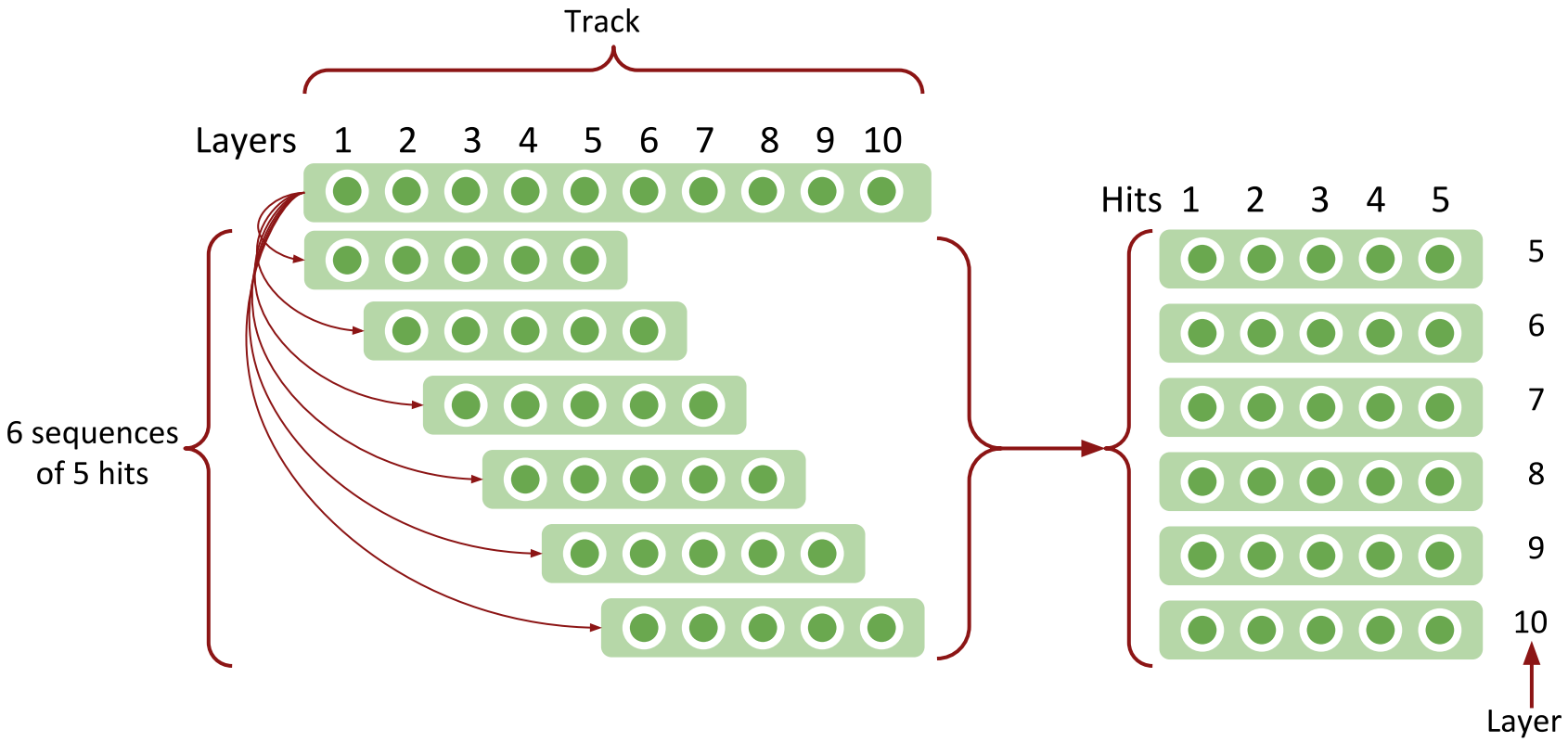


Training Step: GBM Model

```
gbm.x <- gbm( x_5 ~ x_1 + x_2 + x_3 + x_4 +
                    y_1 + y_2 + y_3 + y_4 +
                    z_1 + z_2 + z_3 + z_4,
              data = input_dataset,
              distribution = "gaussian",
              interaction.depth = 20,
              shrinkage = 0.04,
              n.trees = 1000 )
```



Training Step: Preparation of Input Data Set



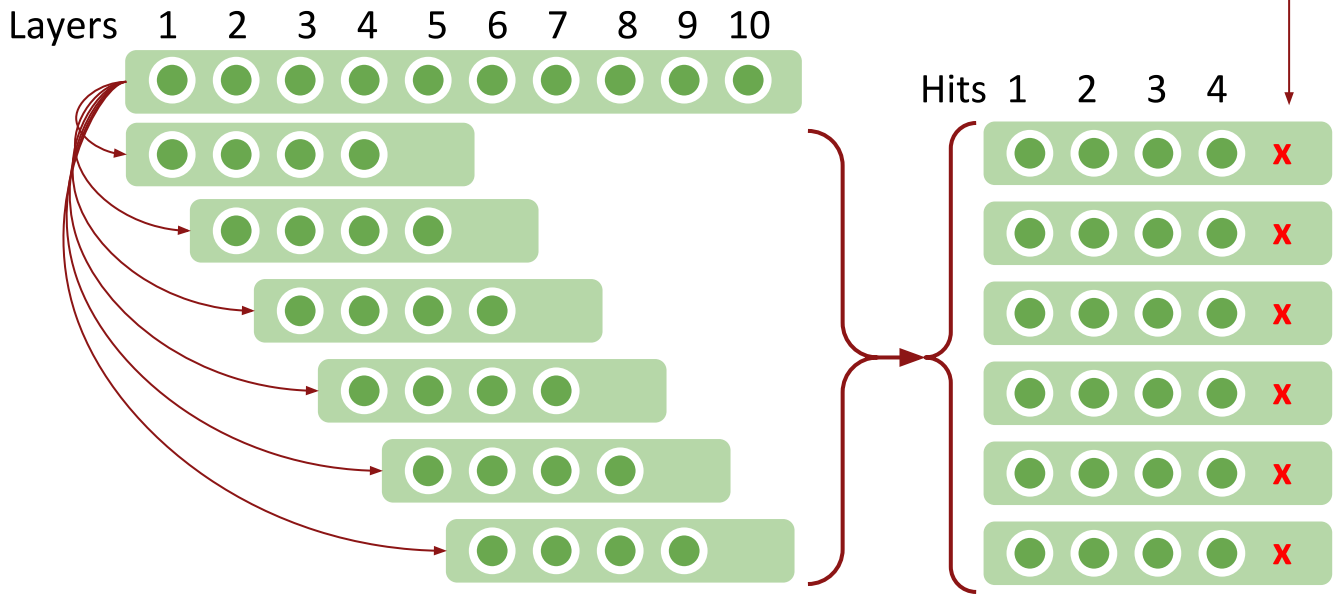
Predicting Coordinates: Input Data Set

```
predict.x <- predict( gbm.x,
              newdata = input_dataset,
              n.trees = 360 )
```

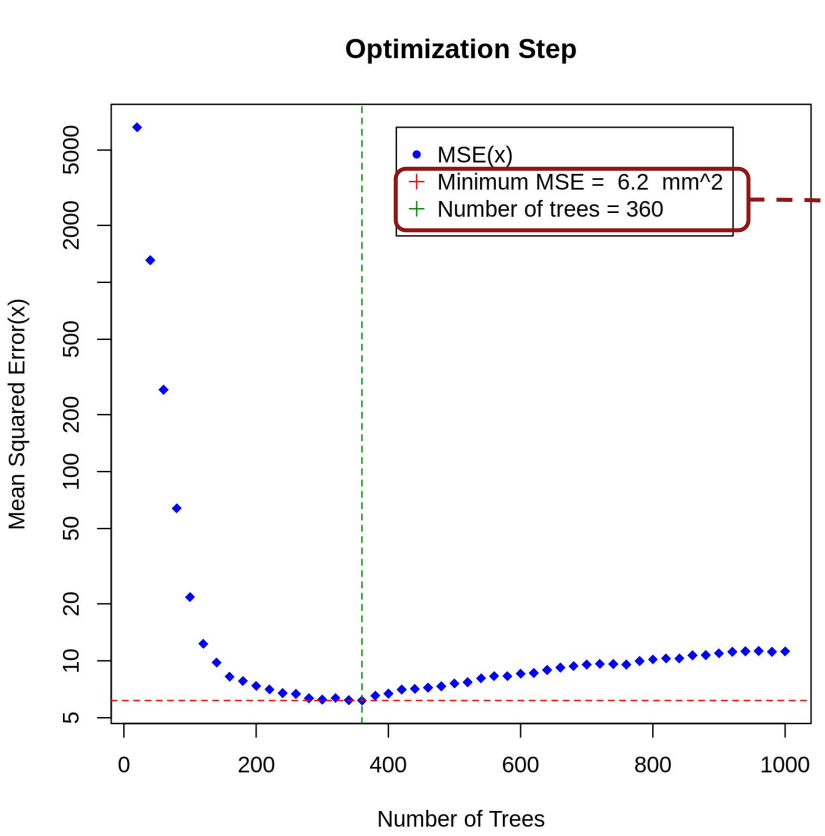
Computed in the Training Step

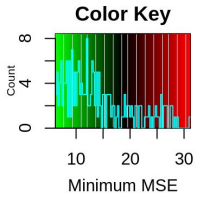
3k tracks different than those for training

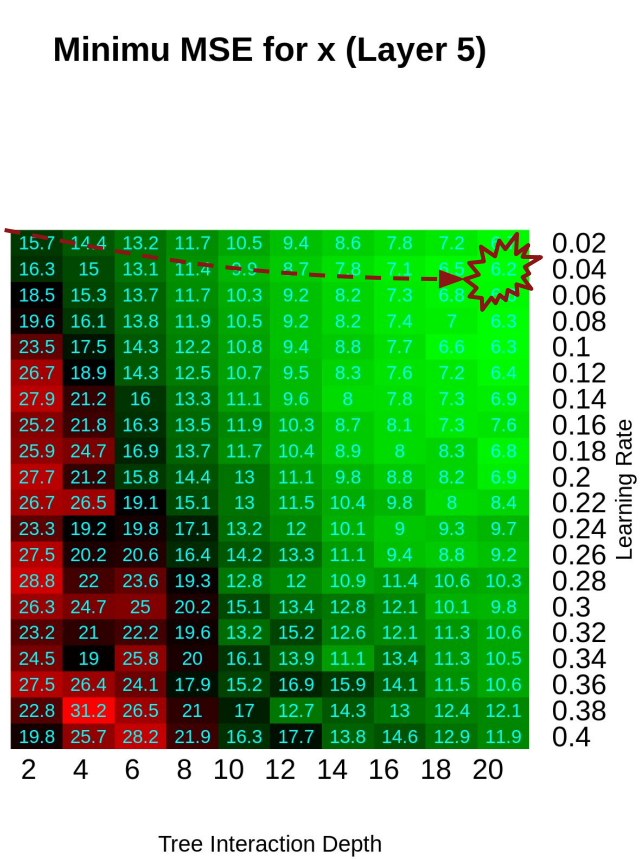
Predicted



Parameters for X (Layer 5)







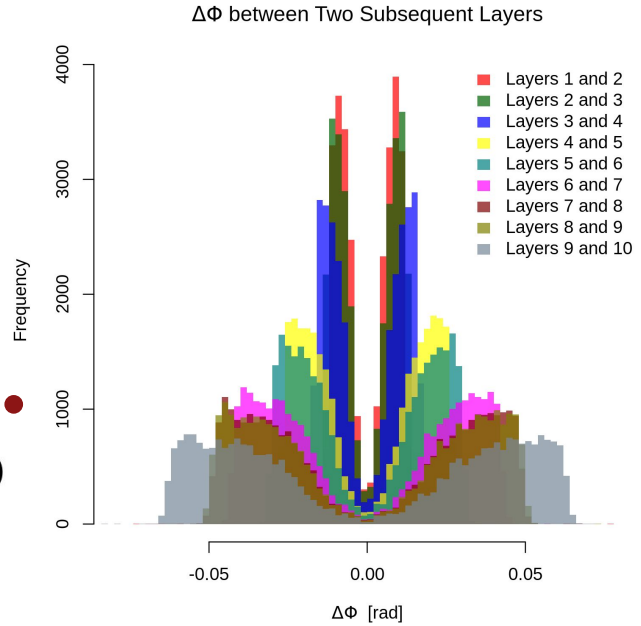
Details about Restrictions

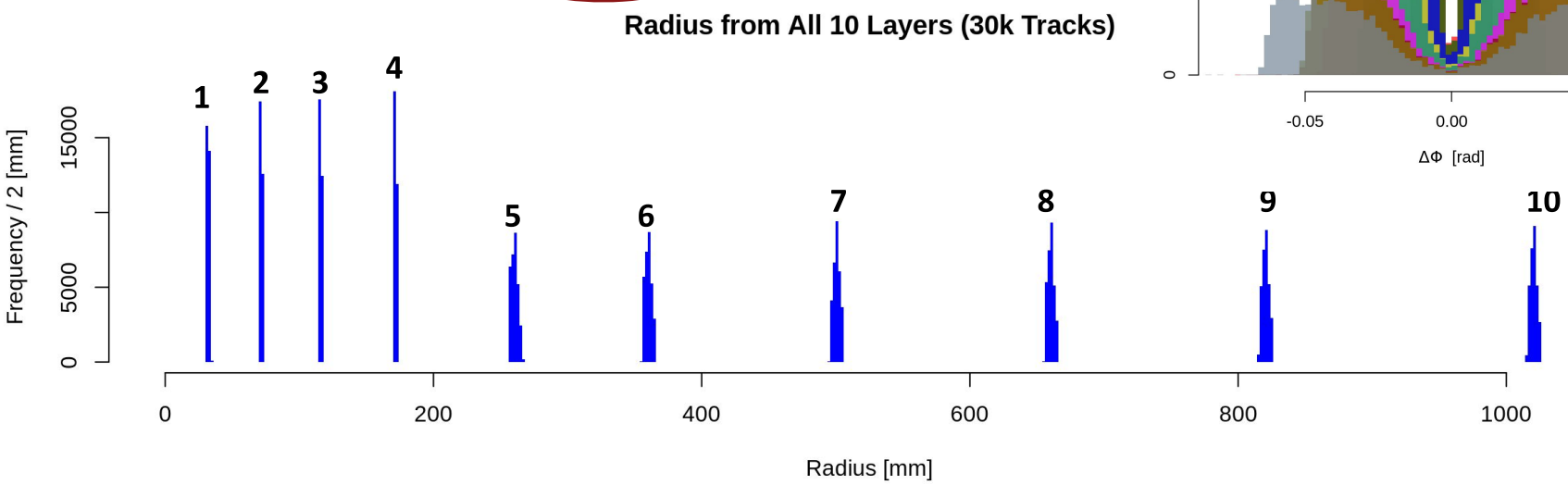
Layer	Volume ID	Layer ID
5	13	2
6	13	4
7	13	6
8	13	8
9	17	2
10	17	4

Performance	No Restrictions	With Restrictions	
Using CPU	Dataset Update	Dataset Update	No Dataset Update
# of Tracks	6000	6000	6000
# of Layers	6	6	6
Total # of Hits	36,000	36,000	36,000
Total Time (s)	1,800	÷ 12 → 150	90
Time to Reconstruct a Single Track (s)	300	0.025	÷ 20 → 0.015
Time to Find Each Hit (s)	50	0.0042	0.0025

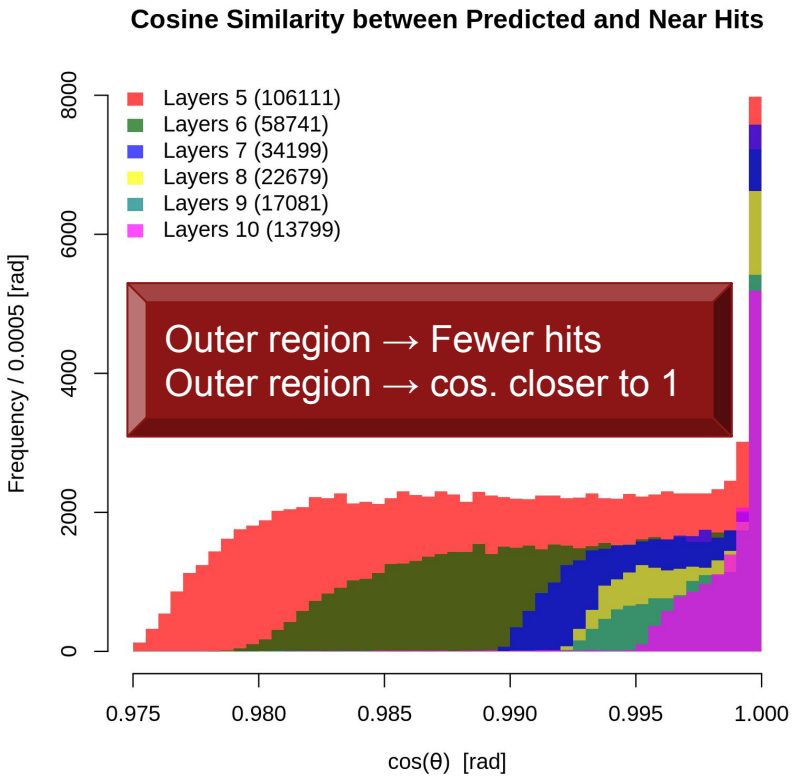
Sanity Check

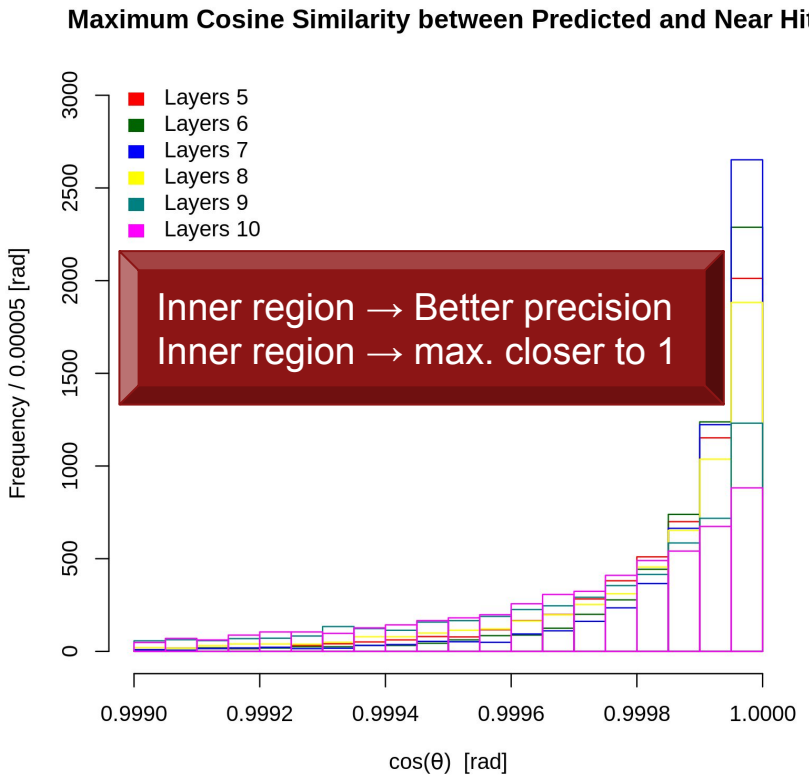
The detector configuration is responsible for ΔΦ distributions. So it is not possible to avoid the track curvature in cosine similarity calculation.

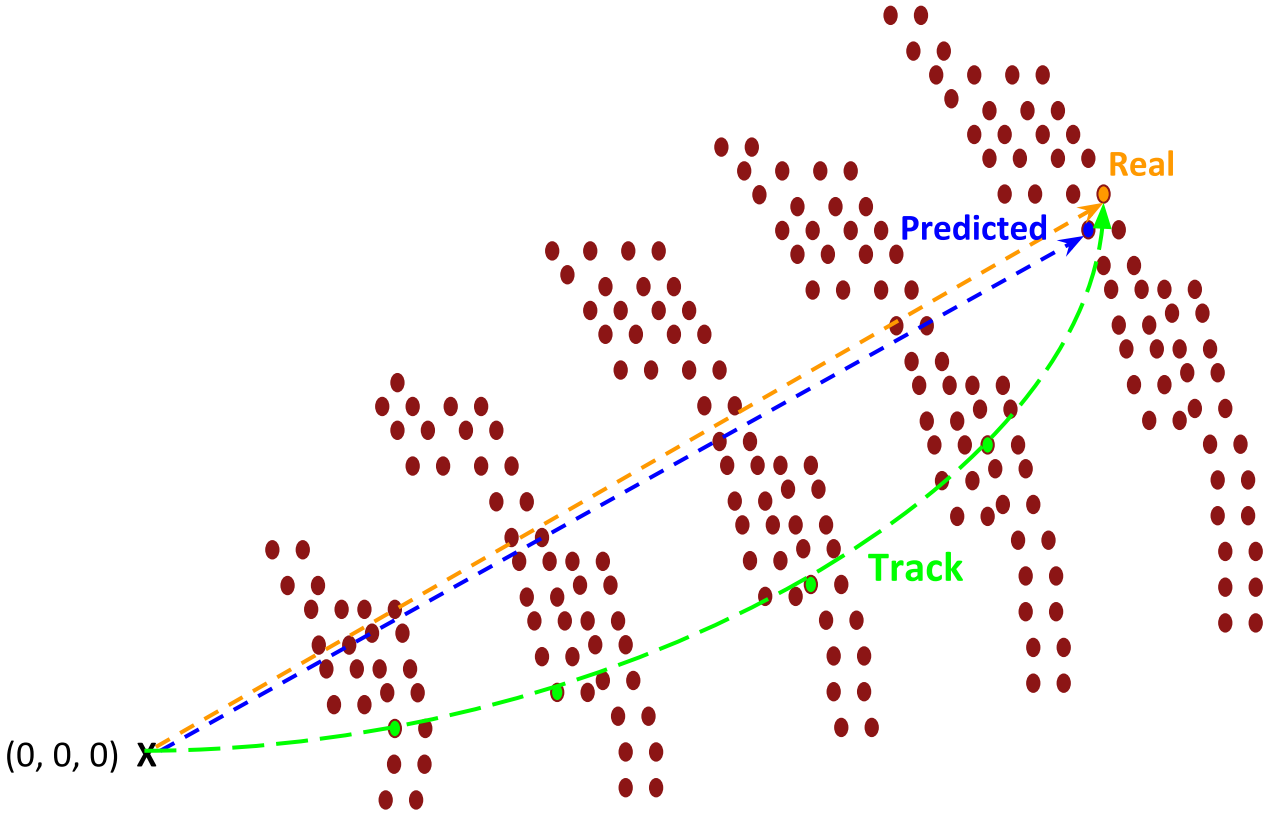




Cosine Similarity Distribution in Each Layer






Real
Predicted
Track
(0, 0, 0)