Beyond the Single Best Model

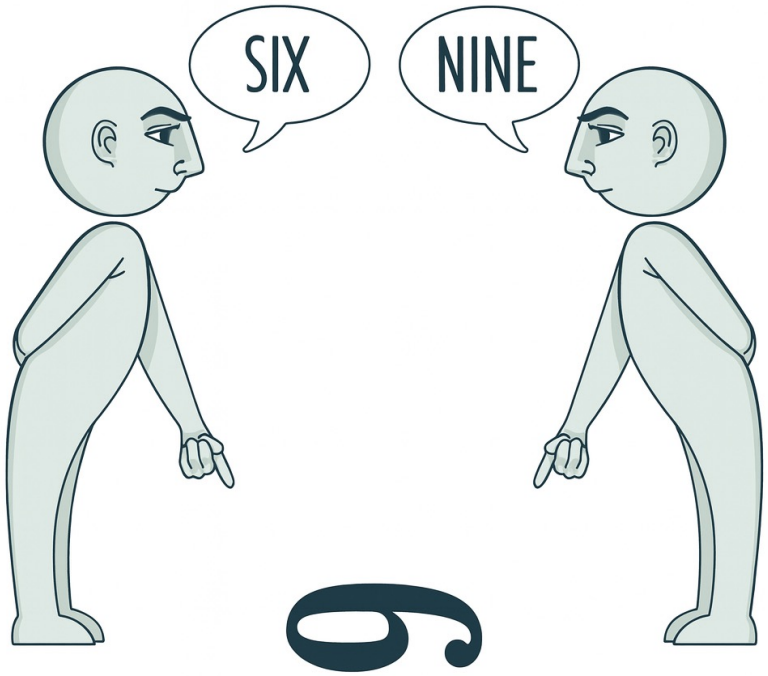
Lesia Semenova

Assistant Professor in CS @ Rutgers

Safe & Transparent AI

How do we make AI systems we can truly understand, trust — and rely on?

No single, “best” path/answer



Rashomon Effect

Many perspectives, all consistent with the same evidence (data).

EPILEPSY

HEALTHY

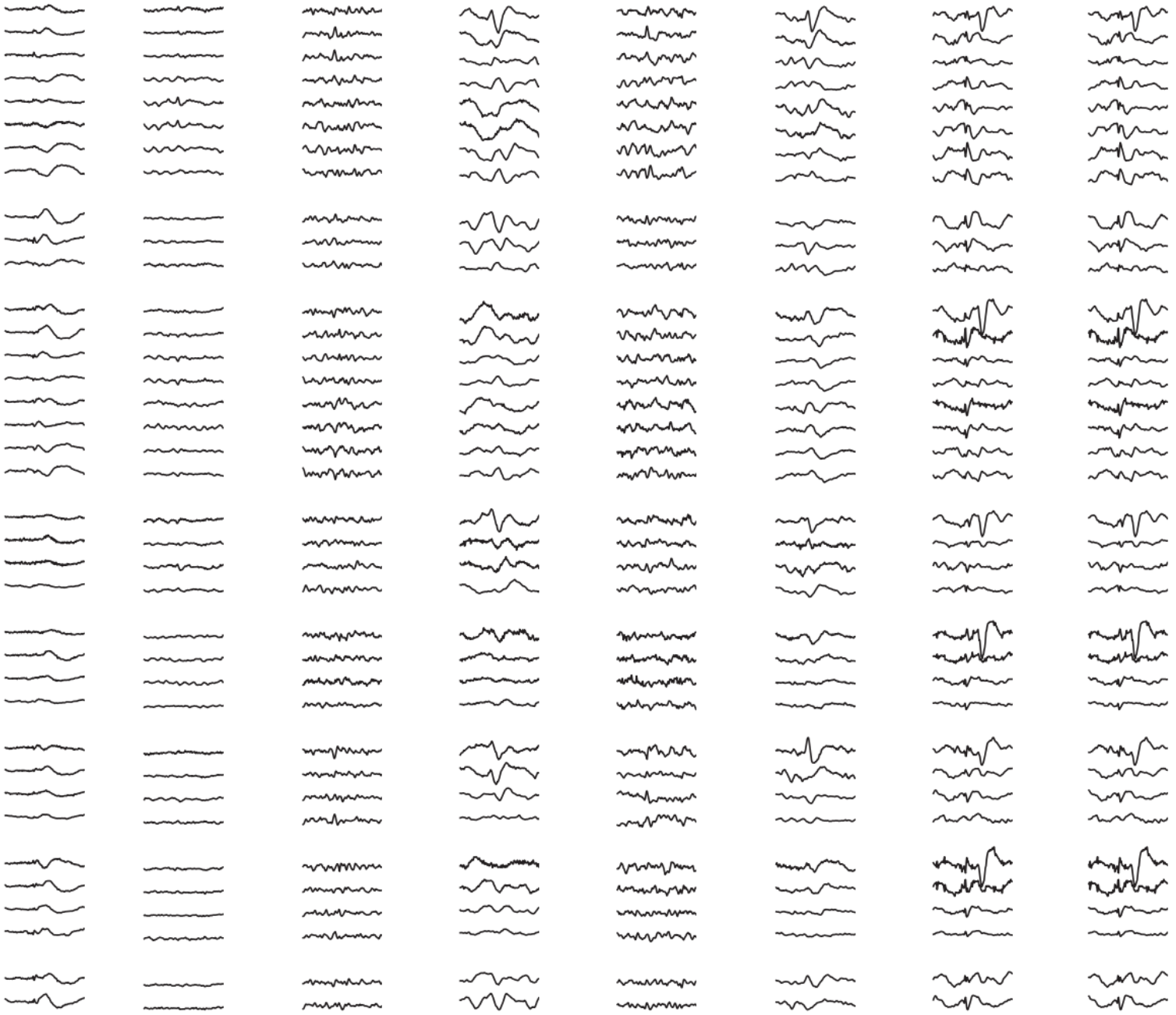
FOCAL
SEIZURE

GENERALIZED
EPLIEPSY

To diagnose epilepsy,

clinicians look **for discharges**

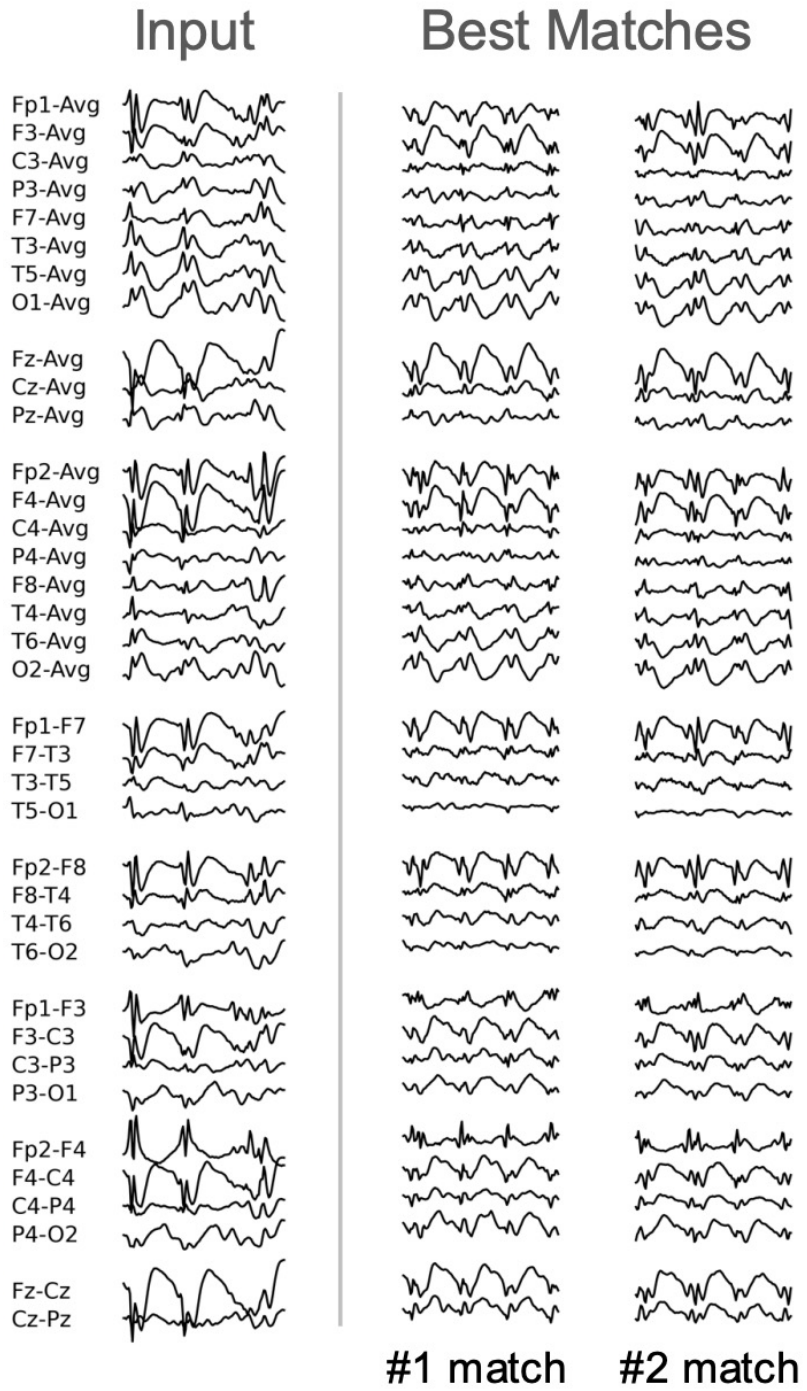
(IEDs) in EEG recordings



Tang et al. MICCAI 2025

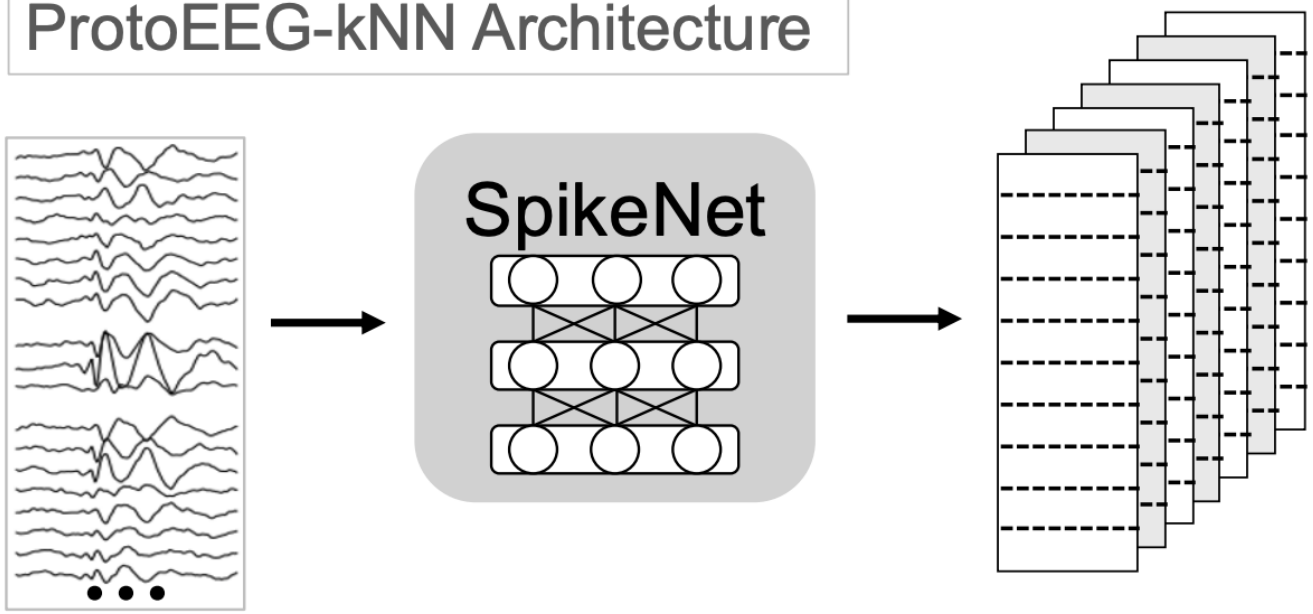
Interpretability: Models that are
inherently constrained
so that their reasoning is understandable
to humans

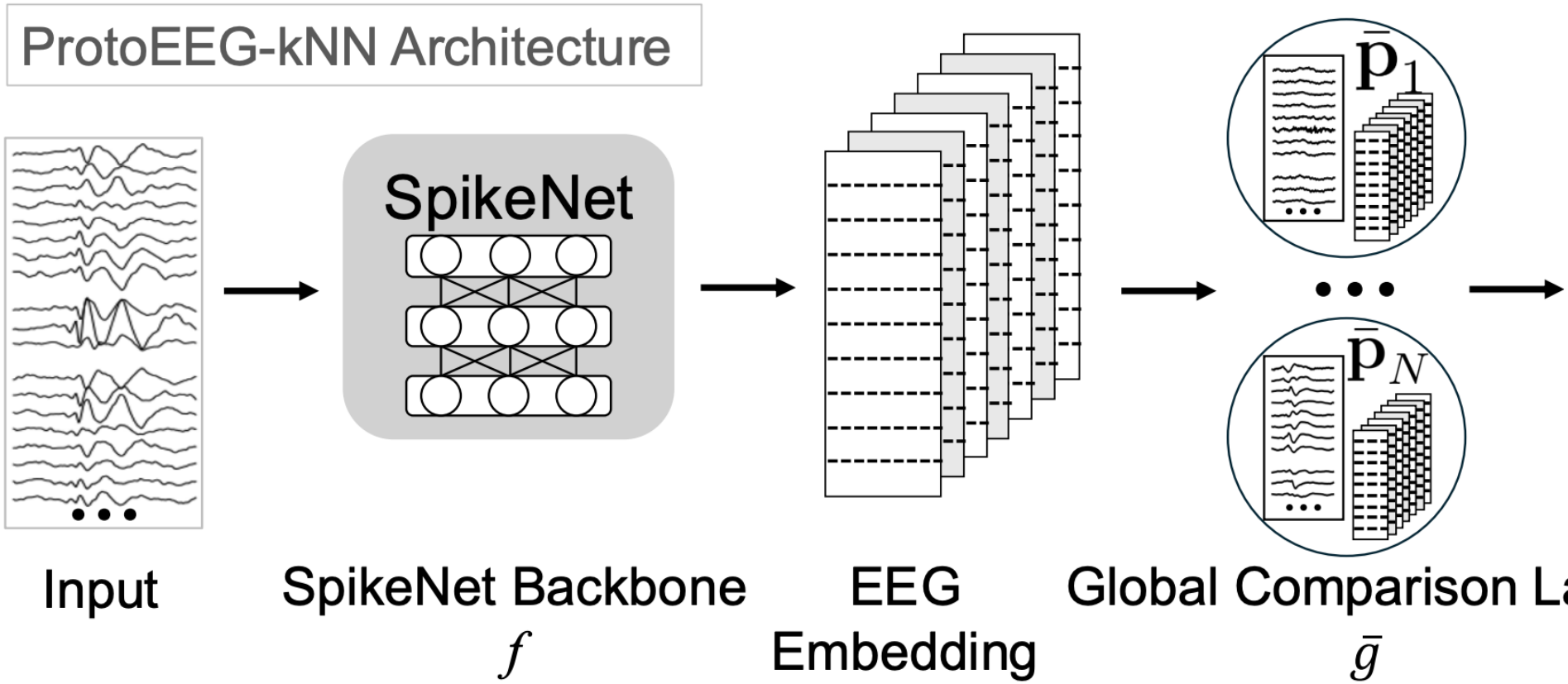
This looks like that
reasoning

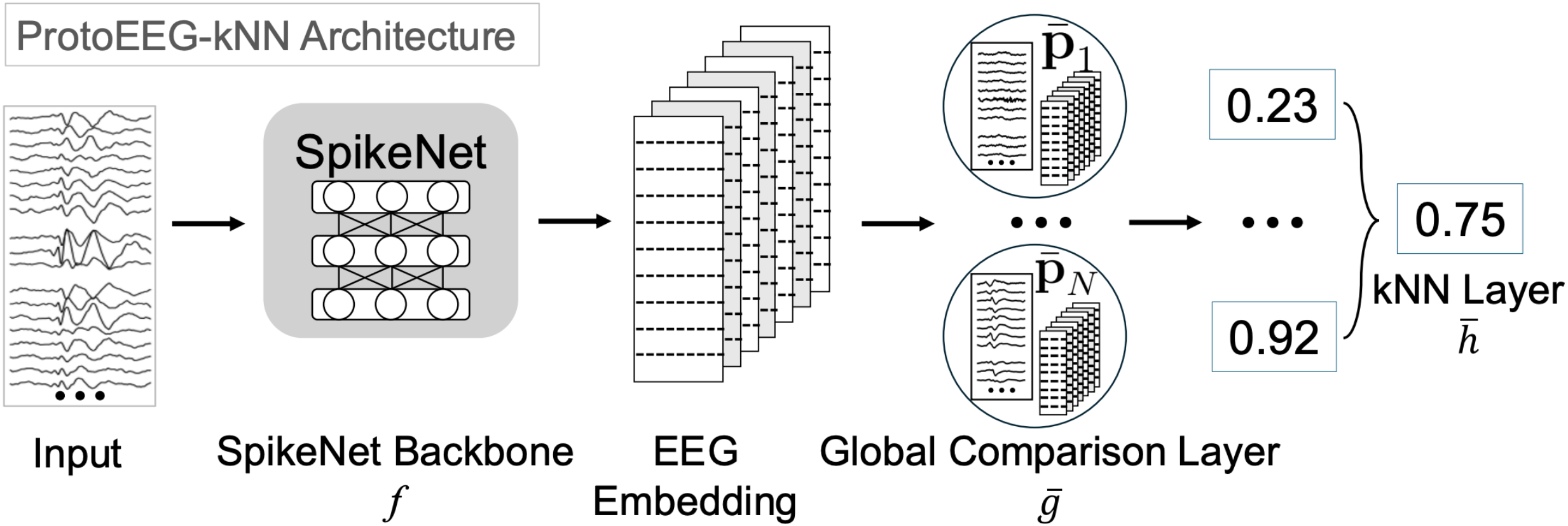


Tang et al. MICCAI 2025

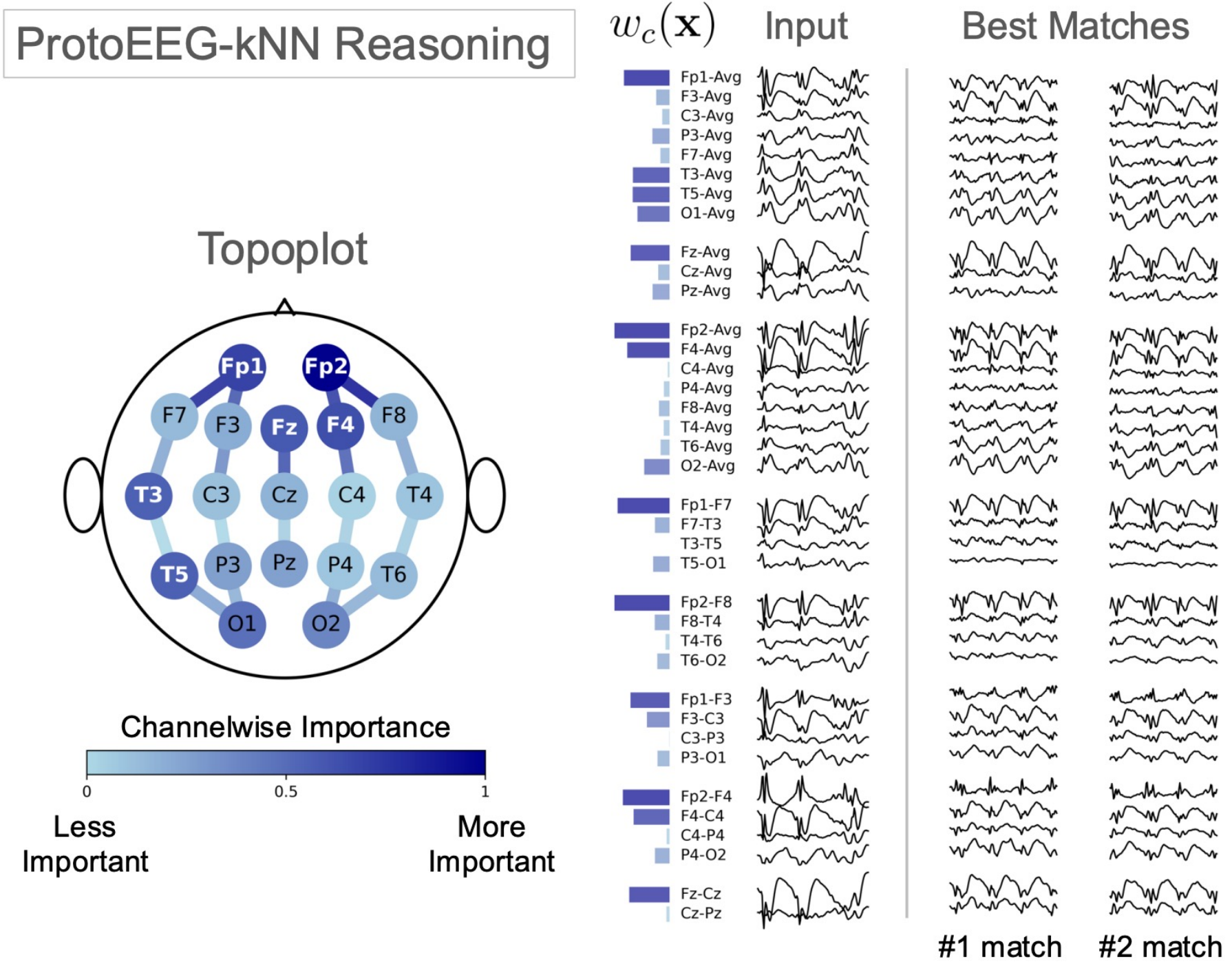
ProtoEEG-kNN Architecture



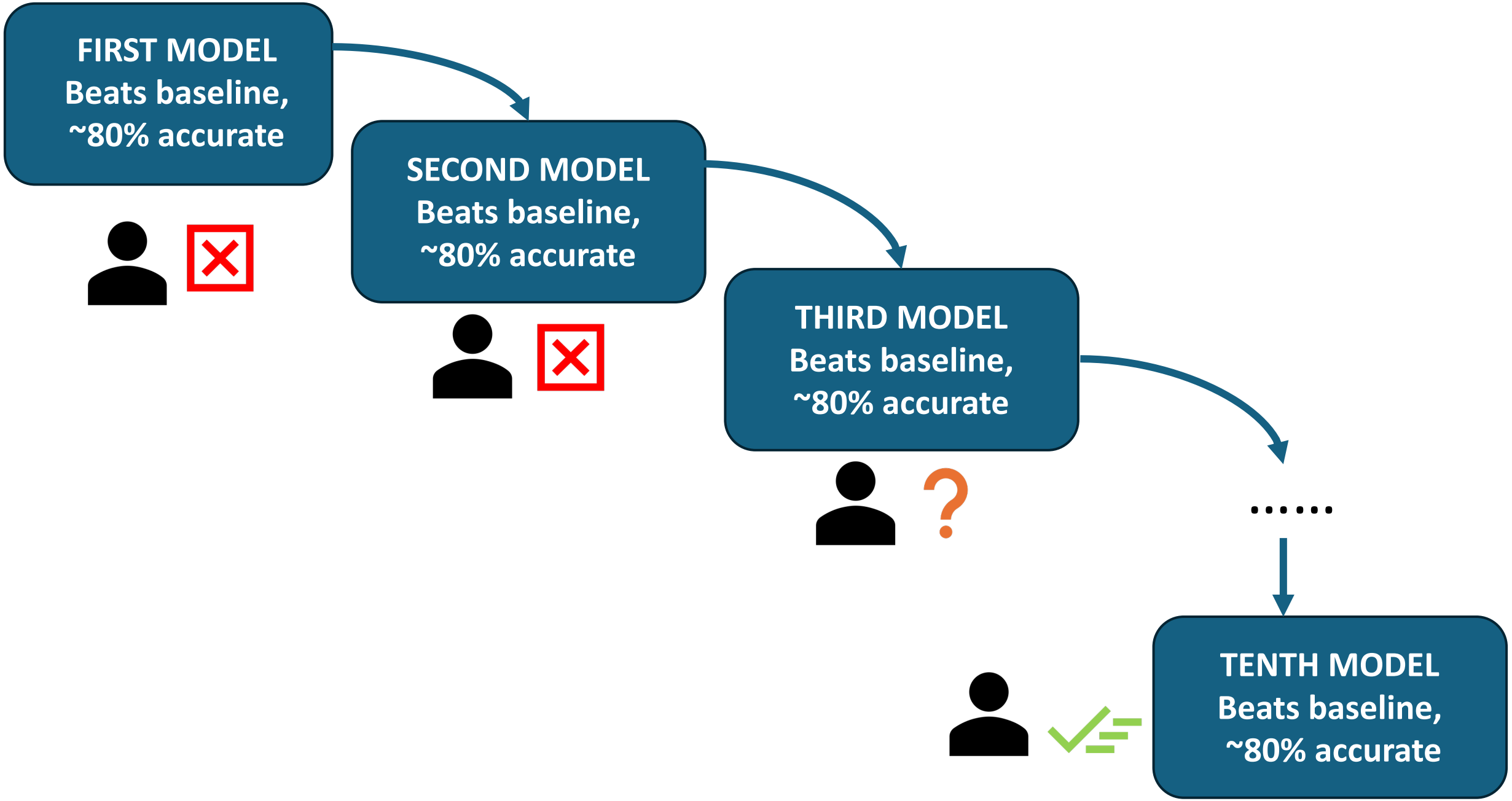

ProtoEEG-kNN Architecture
SpikeNet
$\bar{\mathbf{p}}_1$
$\bar{\mathbf{p}}_N$
Input
SpikeNet Backbone
f
EEG
Embedding
Global Comparison Layer
$\bar{g}$

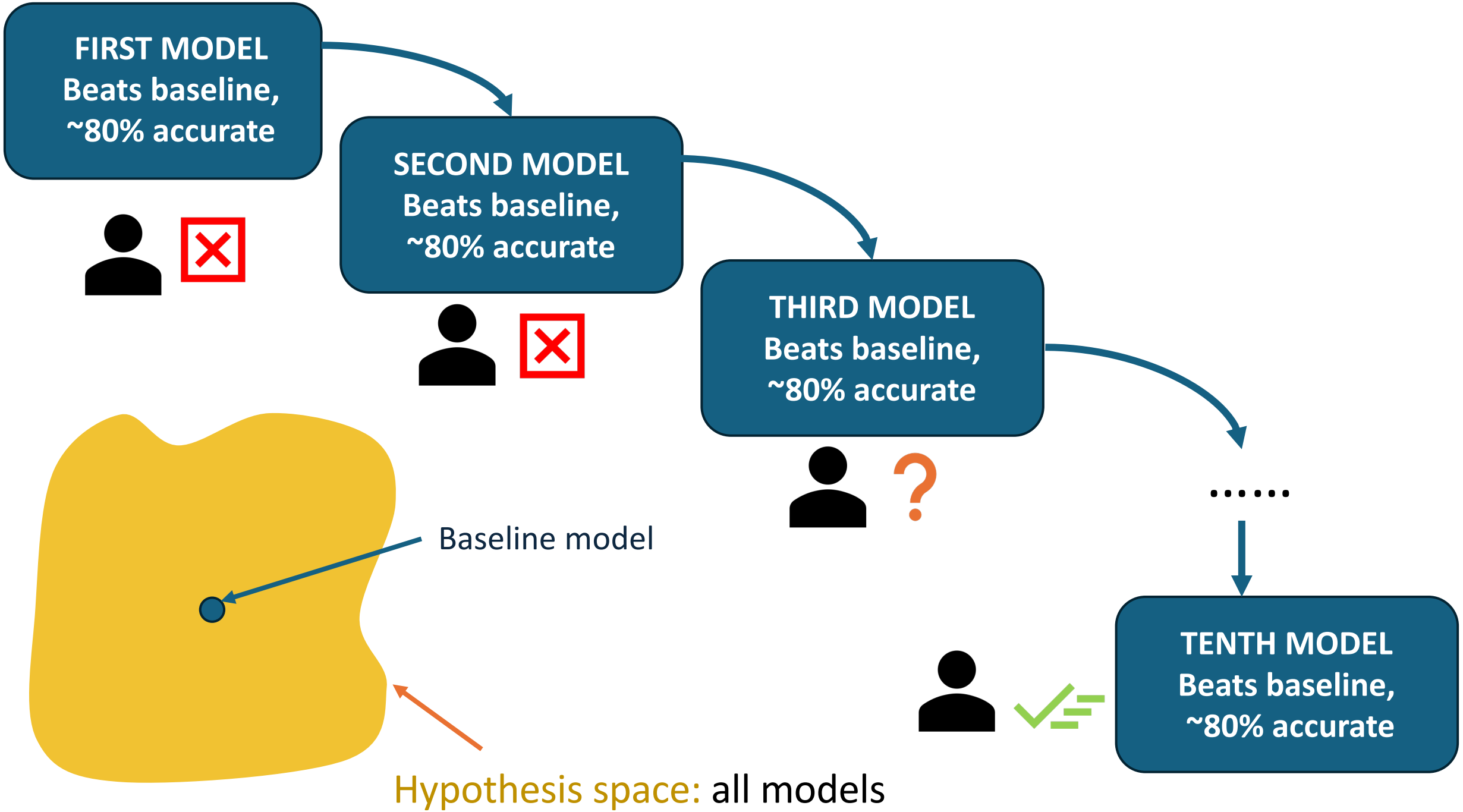

ProtoEEG-kNN Architecture
SpikeNet
$\bar{p}_1$
$\bar{p}_N$
0.23
0.92
0.75
kNN Layer
$\bar{h}$
Input
SpikeNet Backbone
f
EEG
Embedding
Global Comparison Layer
$\bar{g}$

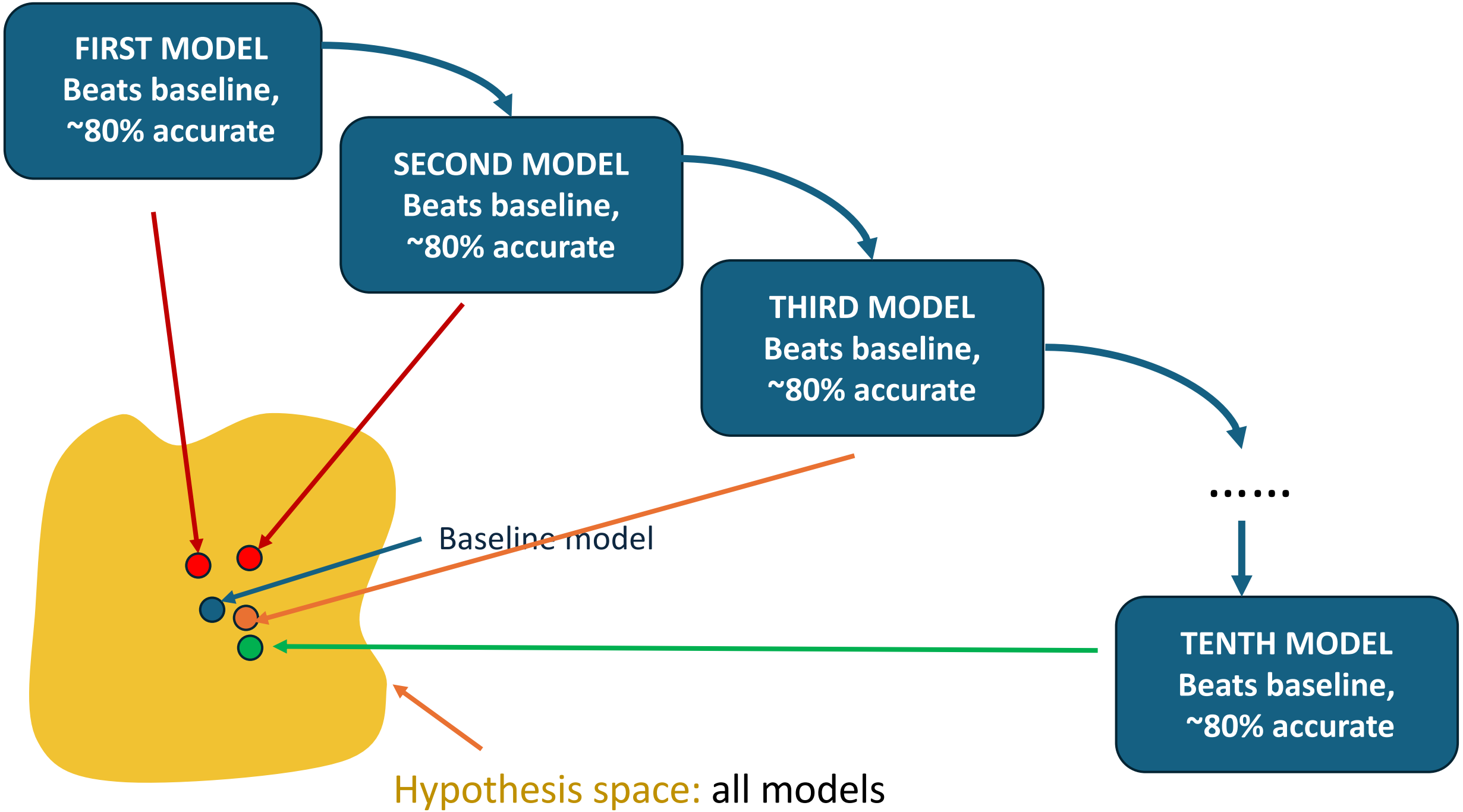
This looks like that reasoning



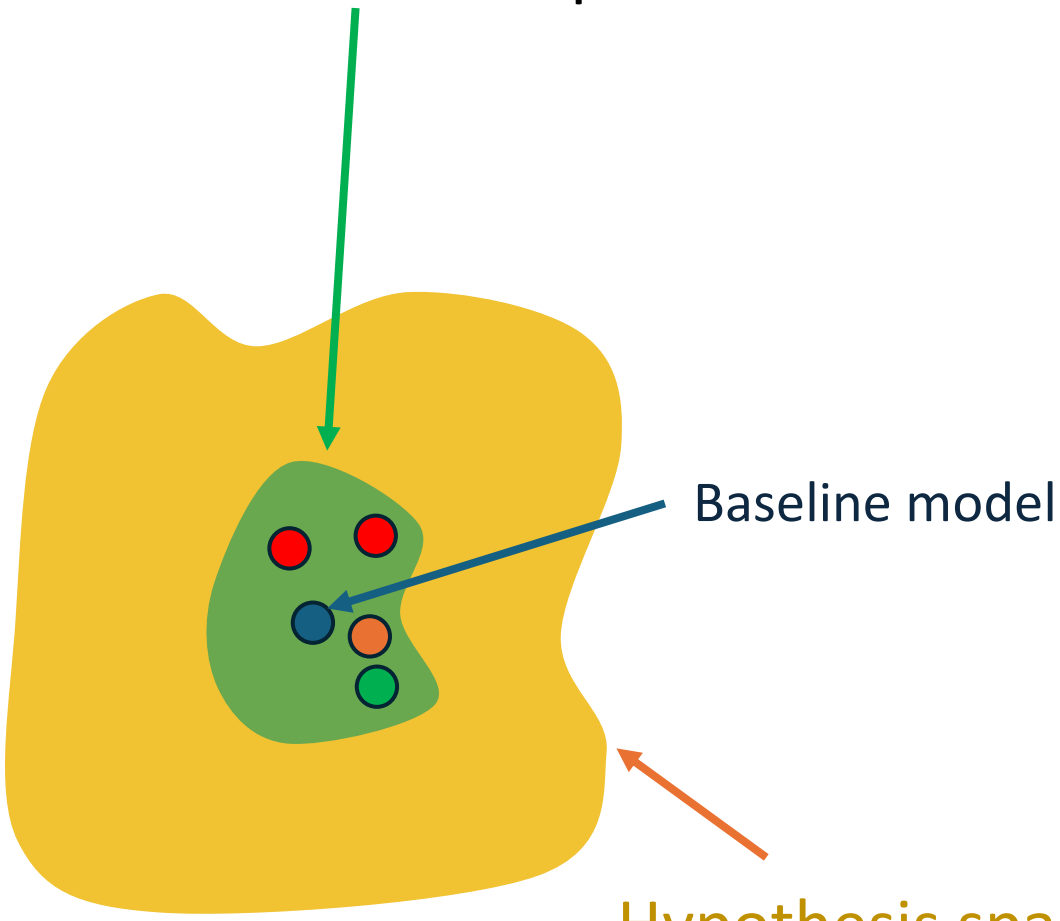
Tang et al. MICCAI 2025


FIRST MODEL
Beats baseline,
~80% accurate
SECOND MODEL
Beats baseline,
~80% accurate
THIRD MODEL
Beats baseline,
~80% accurate
.....
TENTH MODEL
Beats baseline,
~80% accurate


FIRST MODEL
Beats baseline,
~80% accurate
SECOND MODEL
Beats baseline,
~80% accurate
THIRD MODEL
Beats baseline,
~80% accurate
.....
TENTH MODEL
Beats baseline,
~80% accurate
Baseline model
Hypothesis space: all models

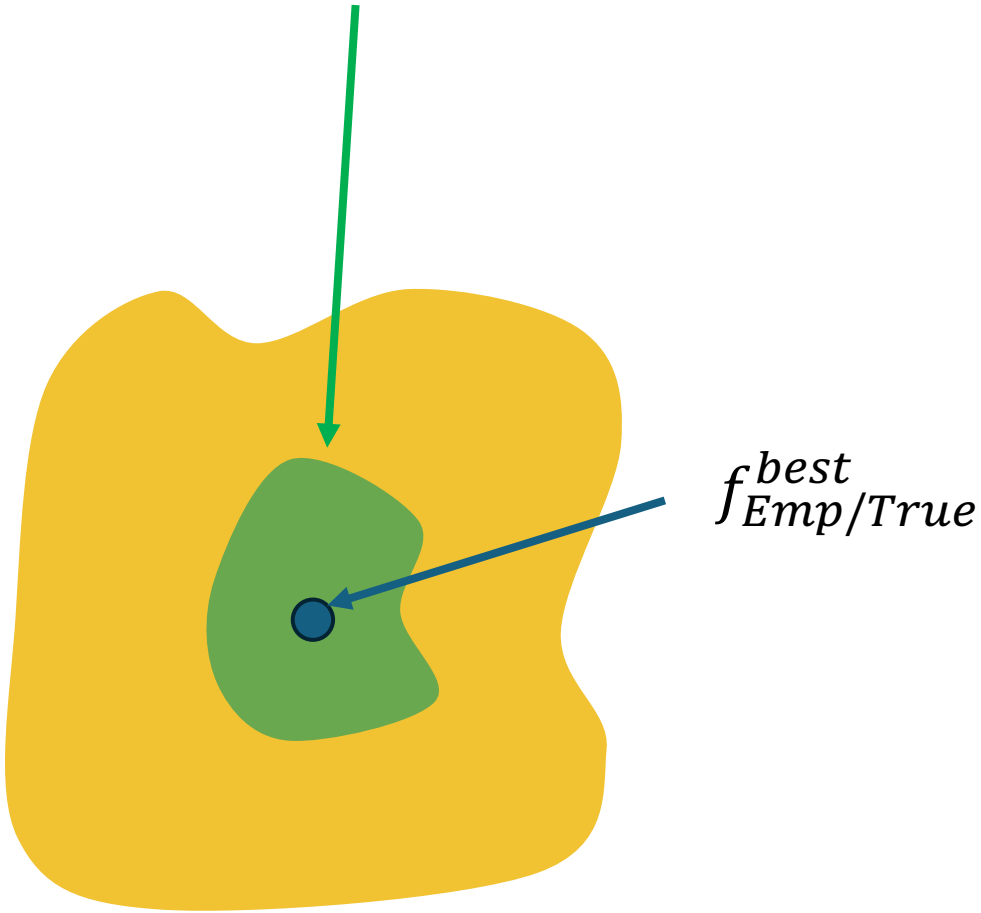

FIRST MODEL
Beats baseline,
~80% accurate
SECOND MODEL
Beats baseline,
~80% accurate
THIRD MODEL
Beats baseline,
~80% accurate
.....
TENTH MODEL
Beats baseline,
~80% accurate
Baseline model
Hypothesis space: all models

set of near-optimal models or **the Rashomon set**



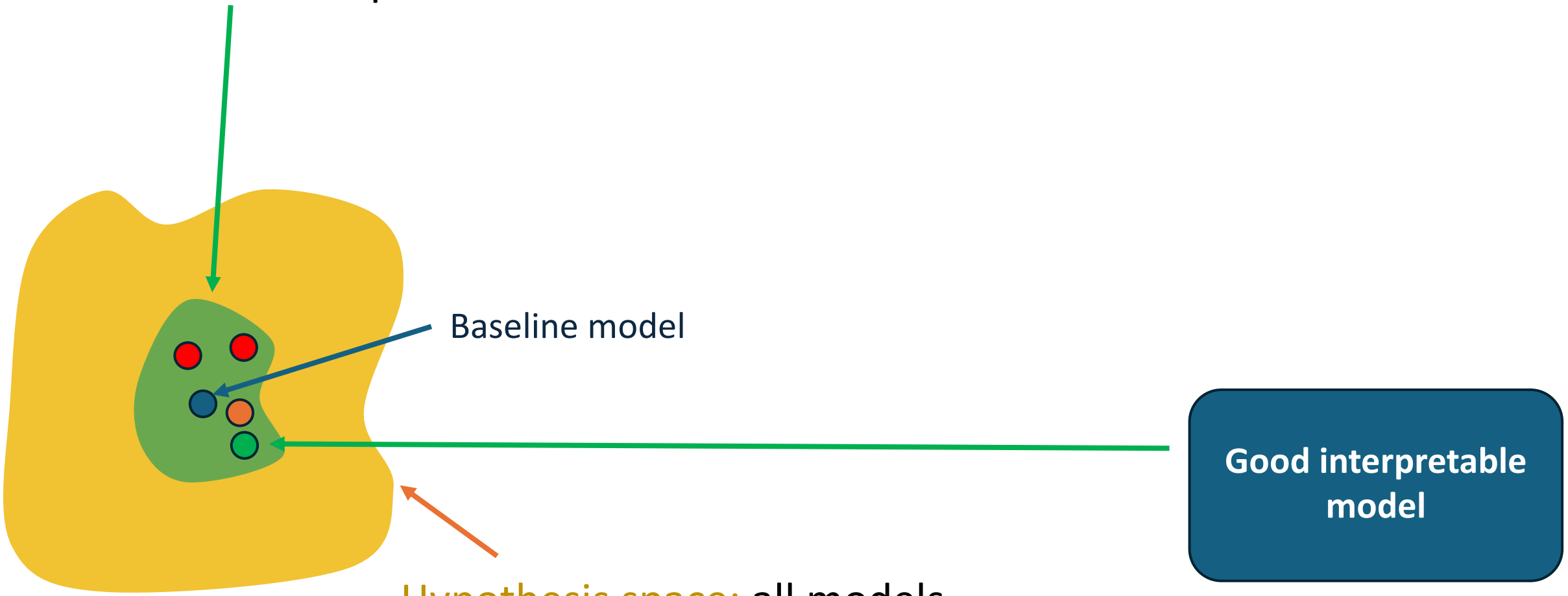
$$R_{set} = \{f : Risk_{Emp/True}(f) \leq Risk_{Emp/True}(f^{best}_{Emp/True}) + \theta\}$$

Rashomon parameter



$f^{best}_{Emp/True}$

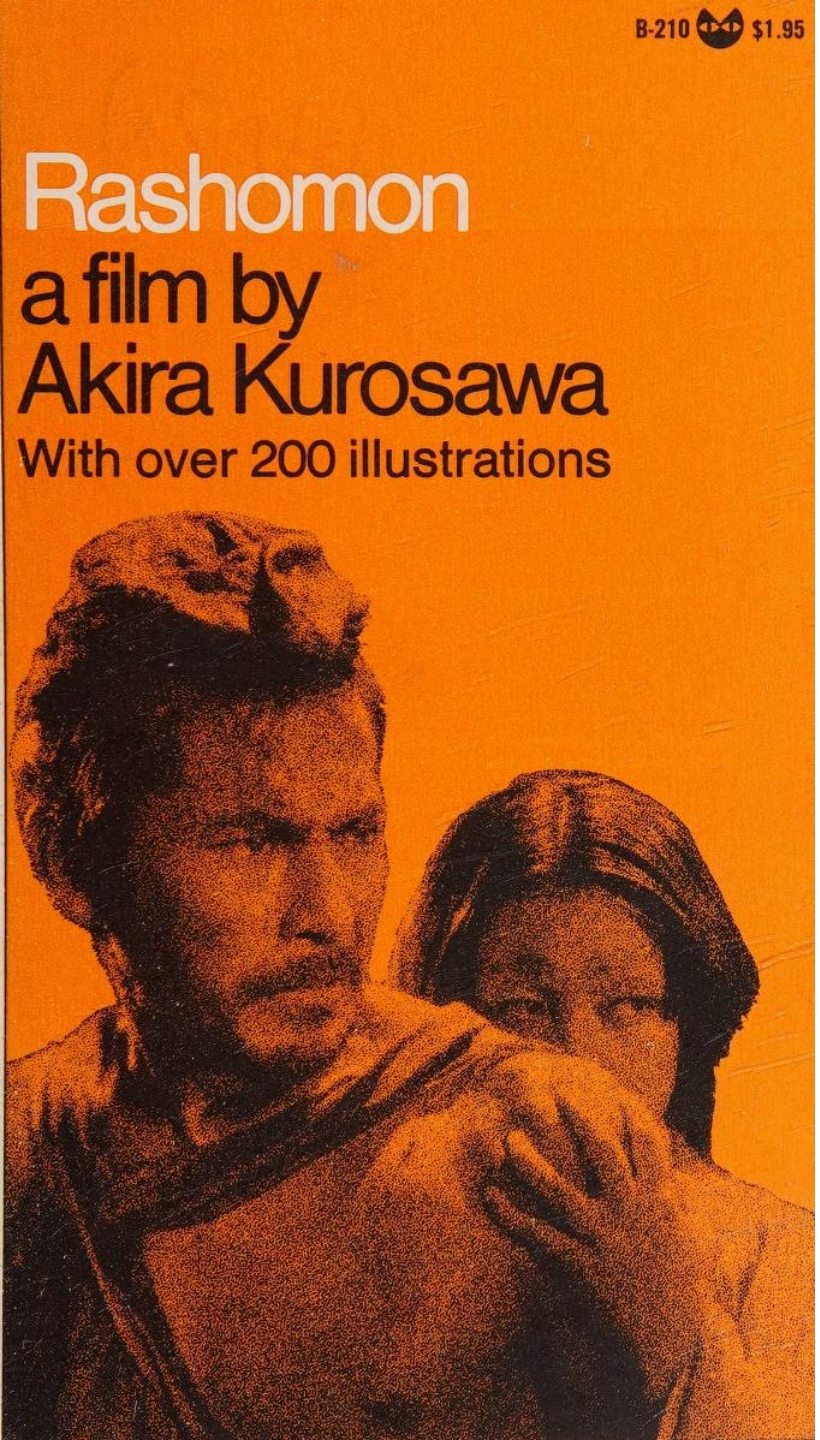
set of near-optimal models or **the Rashomon set**



TAKEAWAY:
If we have large Rashomon set, we can simply

search it for the desirable property

and if the property changes, we can search again (!)

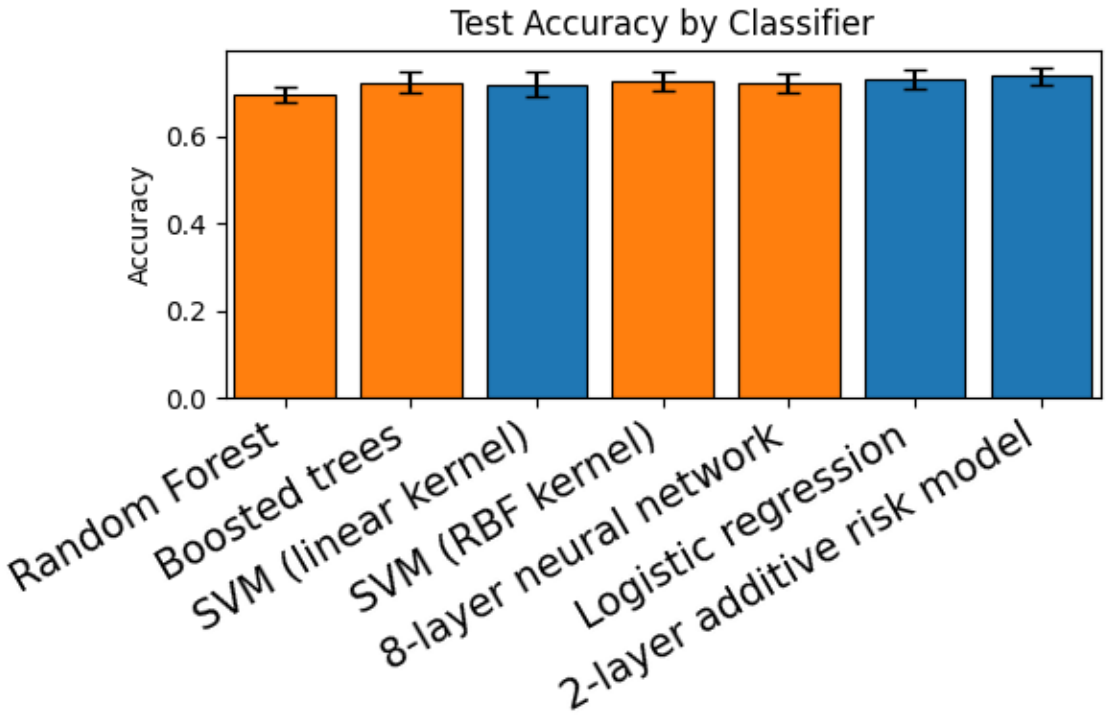


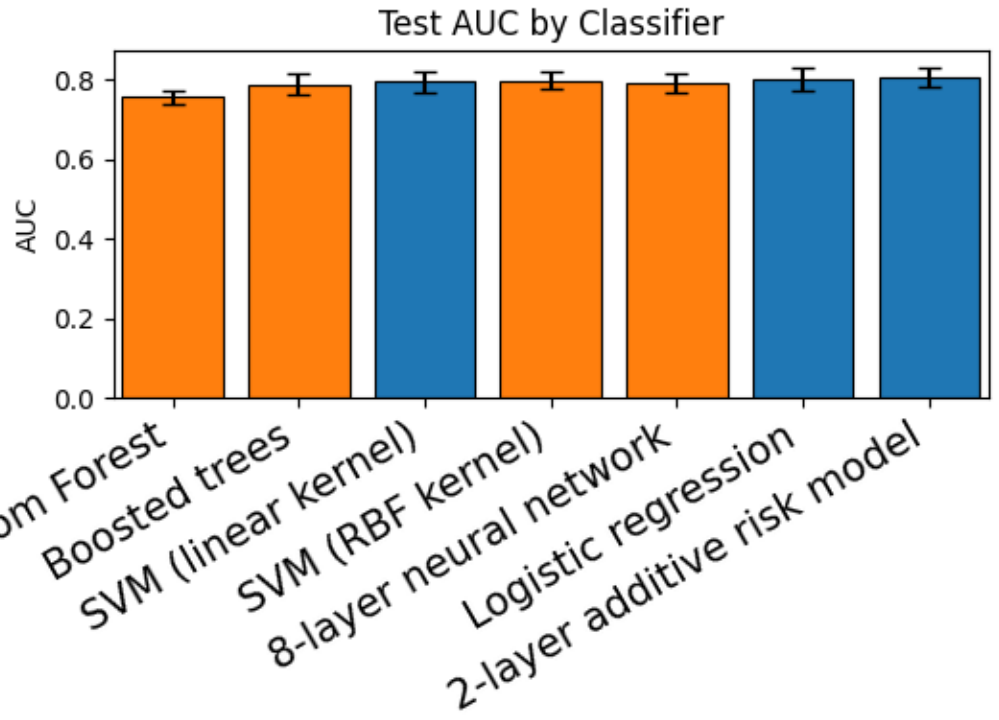
"What I call the **Rashomon Effect** is that there is often a multitude of different descriptions in a class of functions giving about the same minimum error rate."

Leo Breiman (2001)

The Rashomon Effect is everywhere

- Explainable ML Challenge – FICO dataset





Simpler/interpretable models

Black box models

Rudin, Zhong, Semenova et al., ICML 2024

The Rashomon Effect is everywhere

- **Underspecification.** For a give prompts, there could be multiple plausible answers (Hou et al., ICML 2024)

- **Representation Ambiguity.** Two BERT models with different seeds can achieve identical test scores, yet one learns robust linguistic rules while the other relies on fragile keyword-matching shortcuts (McCoy et al., ACL 2020, D'Amour et al., JMLR 2020)

Why does the Rashomon Effect occur?

Underspecification

When a system or learning problem isn't tightly defined by objectives, data, or evaluation, so several models can satisfy it

D'Amour et al., 2020, JMLR

Noise in the data generating processes

Data generation inherent noise due to the randomness of the world

Semenova at al, 2023, NeurIPS; Boner, Chen, Semenova et al., 2024, NeurIPS

Why does the Rashomon Effect occur?

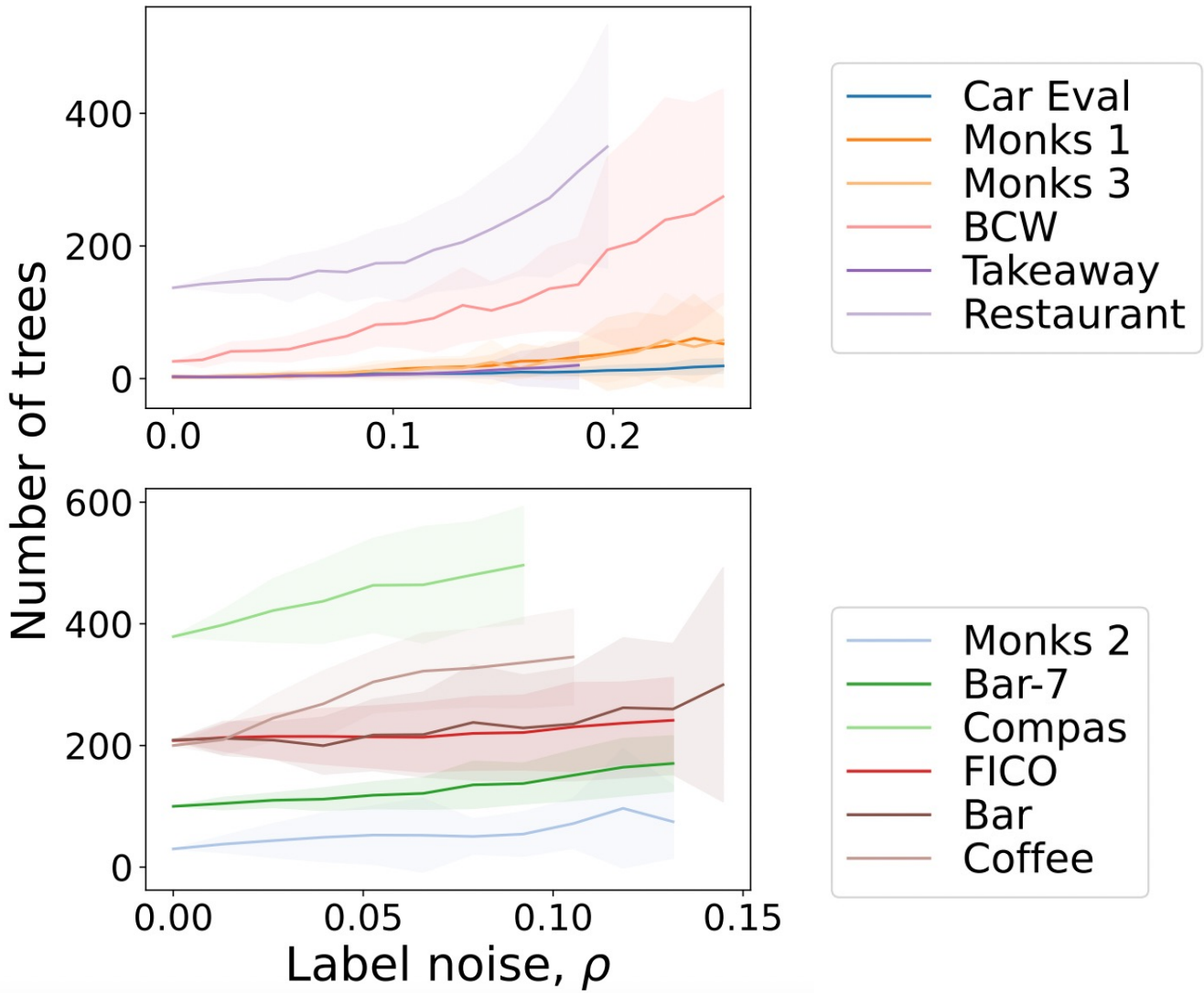
Clean data

Noisy data

Noise is one of the reasons

Noise as one of the causes

The size of the Rashomon set **increases with the random label noise** for the hypothesis space of sparse decision trees



Semenova et al., NeurIPS 2024

Consider random label noise and features g_1 and g_2,
such that $AUC(g_1) < AUC(g_2)$, then

The metric of the better-quality feature decreases faster with noise:

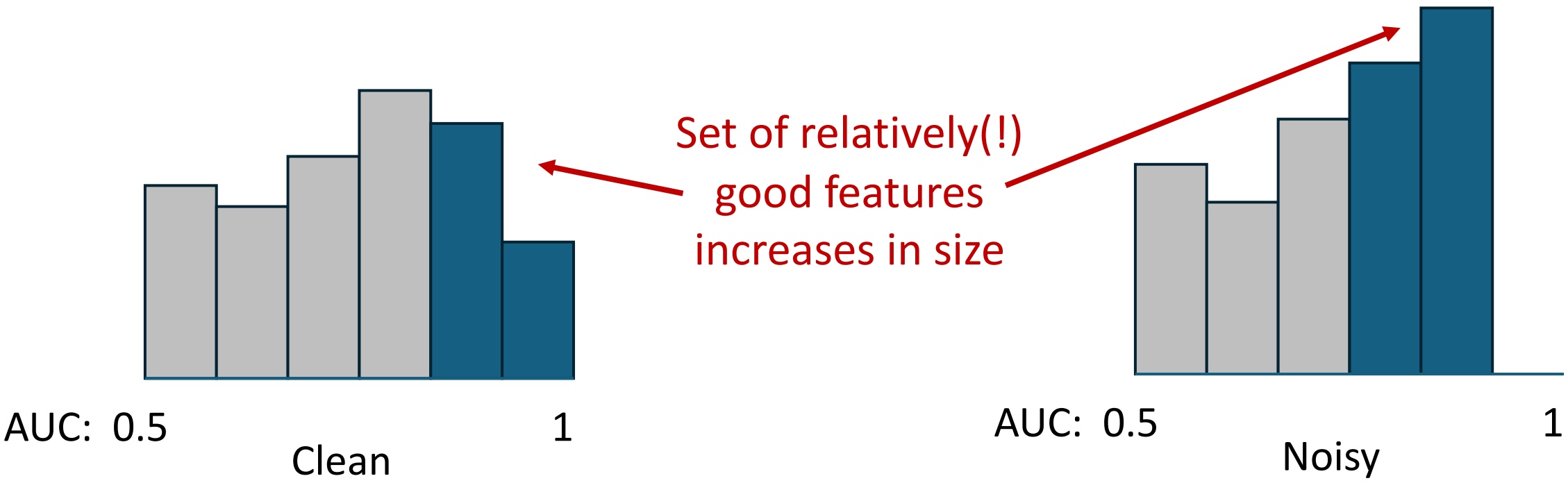
$$\Delta_{noise} AUC(g_1) < \Delta_{noise} AUC(g_2)$$

Theorem 8, Boner, Chen, Semenova, et al., NeurIPS 2024

Consider random label noise and features g_1 and g_2,
such that $AUC(g_1) < AUC(g_2)$, then

The metric of the better-quality feature decreases faster with noise:

$$\Delta_{noise} AUC(g_1) < \Delta_{noise} AUC(g_2)$$



Theorem 8, Boner, Chen, Semenova, et al., NeurIPS 2024

For unregularized hypothesis space of decision trees proved that

the set of models that use relatively good features

increases with noise

Theorem 8, Boner, Chen, Semenova, et al., NeurIPS 2024

For unregularized hypothesis space of decision trees proved that

the set of models that use relatively good features

increases with noise

Informally, noise distorts the feature signal, which allows many different models to achieve similar performance on the same dataset.

Thus, we can **expect larger Rashomon sets.**

Theorem 8, Boner, Chen, Semenova, et al., NeurIPS 2024

Why does the Rashomon Effect occur?

Underspecification

When a system or learning problem isn't tightly defined by objectives, data, or evaluation, so several models can satisfy it

D'Amour et al., 2020, JMLR

Noise in data generating processes

Data generation inherent noise due to the randomness of the world

Semenova at al, 2023, NeurIPS; Boner, Chen, Semenova et al., 2024, NeurIPS

Inherent diversity in data/problem solutions

There could be genuinely multiple ways to solve a problem with 100% accuracy

Feng et al., 2025

Why does the Rashomon Effect occur?

Rashomon Concept Bottleneck Models (CBMs) - predict human-understandable "concepts"

Class	Model 0	Model 1	Model 2	Model 3	Model 4
Tiger 	+ strong (+5.76)	- swims (+3.66)	- tusk (-5.63)	+ fierce (+1.39)	- hooves (+16.77)
	+ meatteeth (+2.12)	- longneck (+1.46)	+ quadrupedal (+3.22)	- grazer (+1.32)	+ claws (+1.38)
	- small (+1.81)	- gray (+1.38)	+ fast (+1.95)	- skimmer (+1.02)	- newworld (+1.00)
	- desert (-1.10)	+ orange (+1.24)	- tunnels (-1.53)	- longneck (+0.92)	+ ground (-0.96)
	- domestic (+1.01)	+ straightteeth (+1.13)	- inactive (+1.48)	- vegetation (+0.74)	- plankton (+0.79)

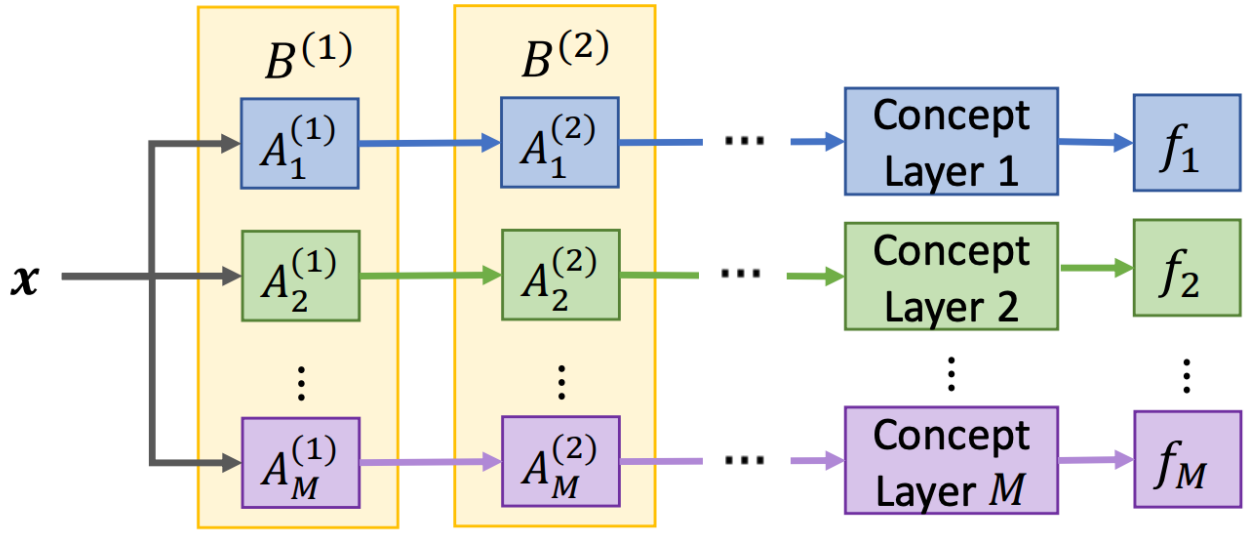
- green - positive evidence
- red - negative evidence
- blue - excluding evidence
- purple - spurious evidence

Feng, Cheng, Xi, Semenova, Zhong, 2025

Rashomon CBMs

Models from the Rashomon set that are different enough per some diversity metric (cos similarity)

min *Task Loss + p1 Concept Loss – p2 Diversity Loss*



(a) Overall structure

an input image passes through a frozen backbone with attached adapters, followed by multiple parallel concept layers, each linked to a classifier.

Feng, Cheng, Xi, Semenova, Zhong, 2025

Methods to compute the Rashomon Effect

- CBMs – Rashomon CBMs - optimizes model diversity (cos distance) + low rank adapters – (Feng, Cheng, Xi, Semenova, Zhong, 2025)
- Prototypical NN - Rashomon ProtoPNets – explores multiplicity of the embedding layer (Donelly et al CVPR 2025)
- NN – changing random seeds, hyperparameter tuning, perturbations (Hsu at al, NeurIPS 2022)
- Trees, generalized additive models – TreeFarms, RashomonGAM – compute exactly or almost exactly (Xin at al, NeurIPS 2022, Zhong at al, NeurIPS 2023)

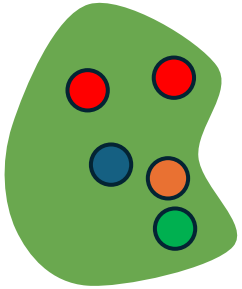
The Rashomon Effect changes almost everything we know in machine learning

The Rashomon Effect changes almost everything we know in machine learning

When **A LOT** of models that perform approximately-equally well

1. Which model to choose?
2. What variables are important?
3. Are we explaining the correct model?
4. What is the problem complexity?
5. Is there a fair model?
6. Does there exist accurate interpretable model?

……



Source of uncertainty, but full of magic and possibilities

Turning Rashomon Effect

from a **source of uncertainty**

into a **foundation for trust**

TAKEAWAY 1:
If we have large Rashomon set, we can simply

search it for the desirable property

and if the property changes, we can search again (!)

Benchmarked Rashomon set on

- Robustness (to adversarial attacks)
- Stability (to noise)
- Privacy (membership-inference attack)
- Fairness (different disparities)

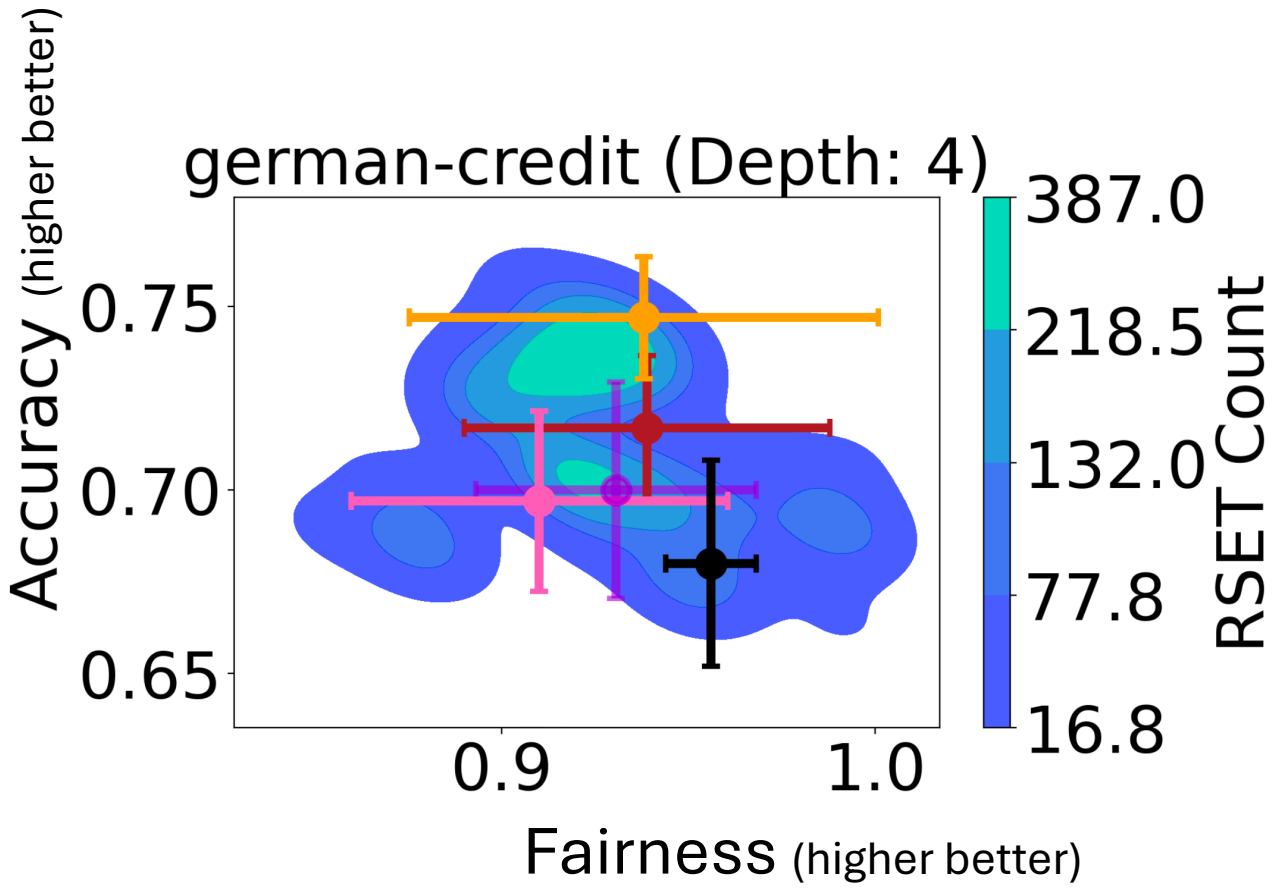
Consider best methods that optimize for the desiderata directly and compare them with models in the Rashomon set

The Rashomon set has it all. Hsu, Cao, Semenova, Zhong, *NeurIPS 2025*

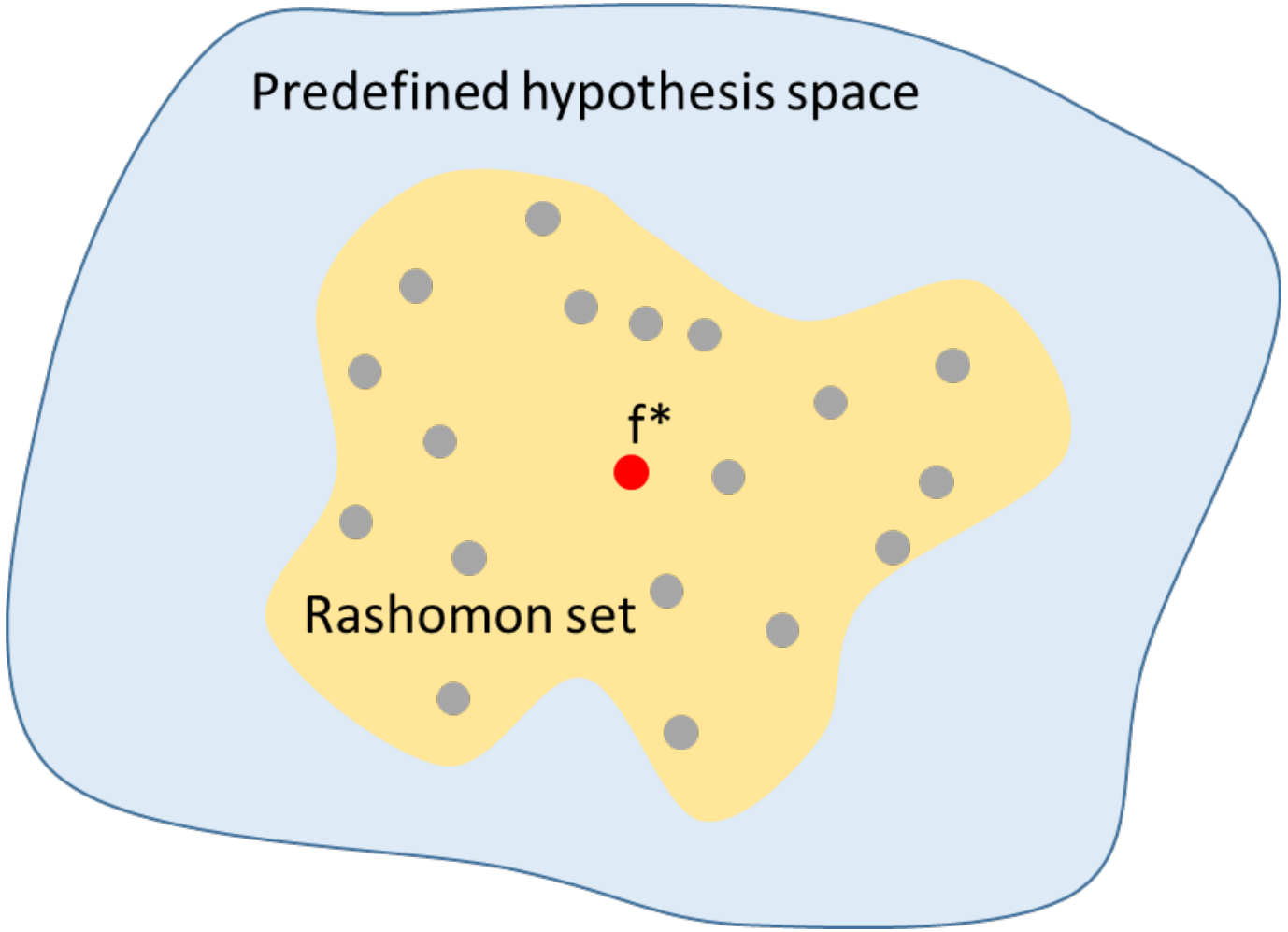
Benchmarked Rashomon set on

- Robustness (to adversarial attacks)
- Stability (to noise)
- Privacy (membership-inference attack)
- Fairness (different disparities)

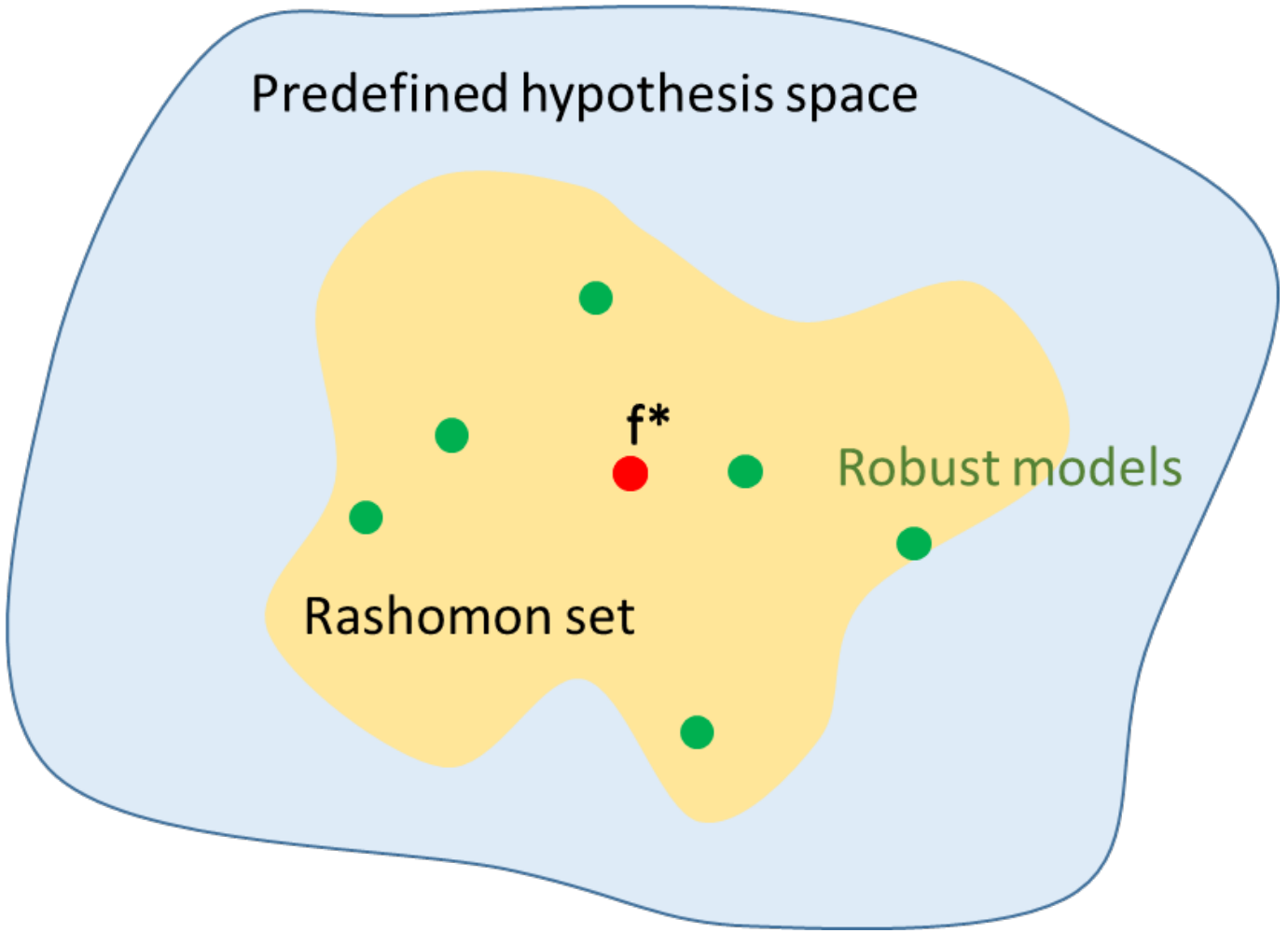
We show that models from the Rashomon set can **perform comparably** to models **explicitly optimized for trustworthiness**



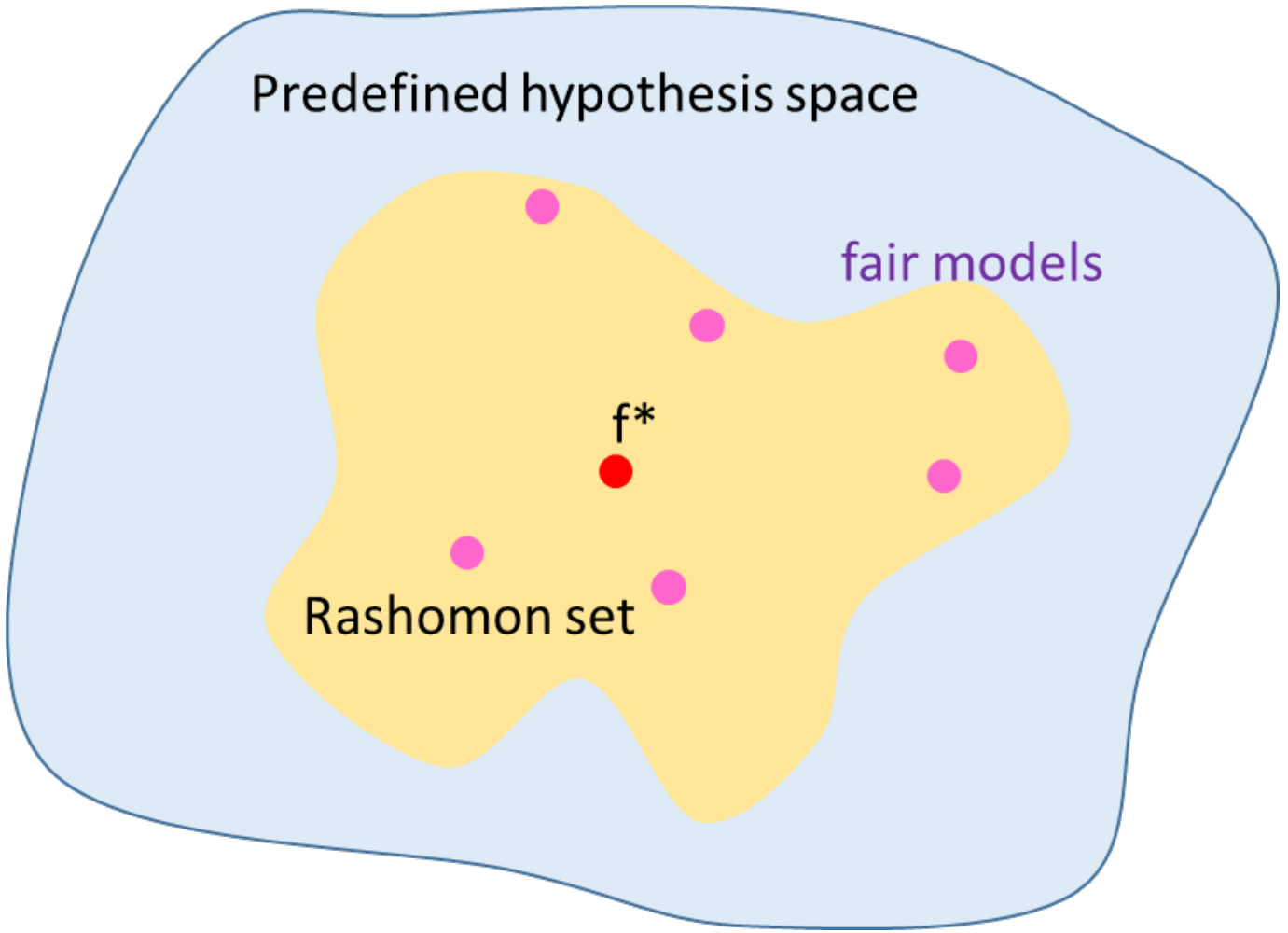
The Rashomon set has it all. Hsu, Cao, Semenova, Zhong, *NeurIPS 2025*



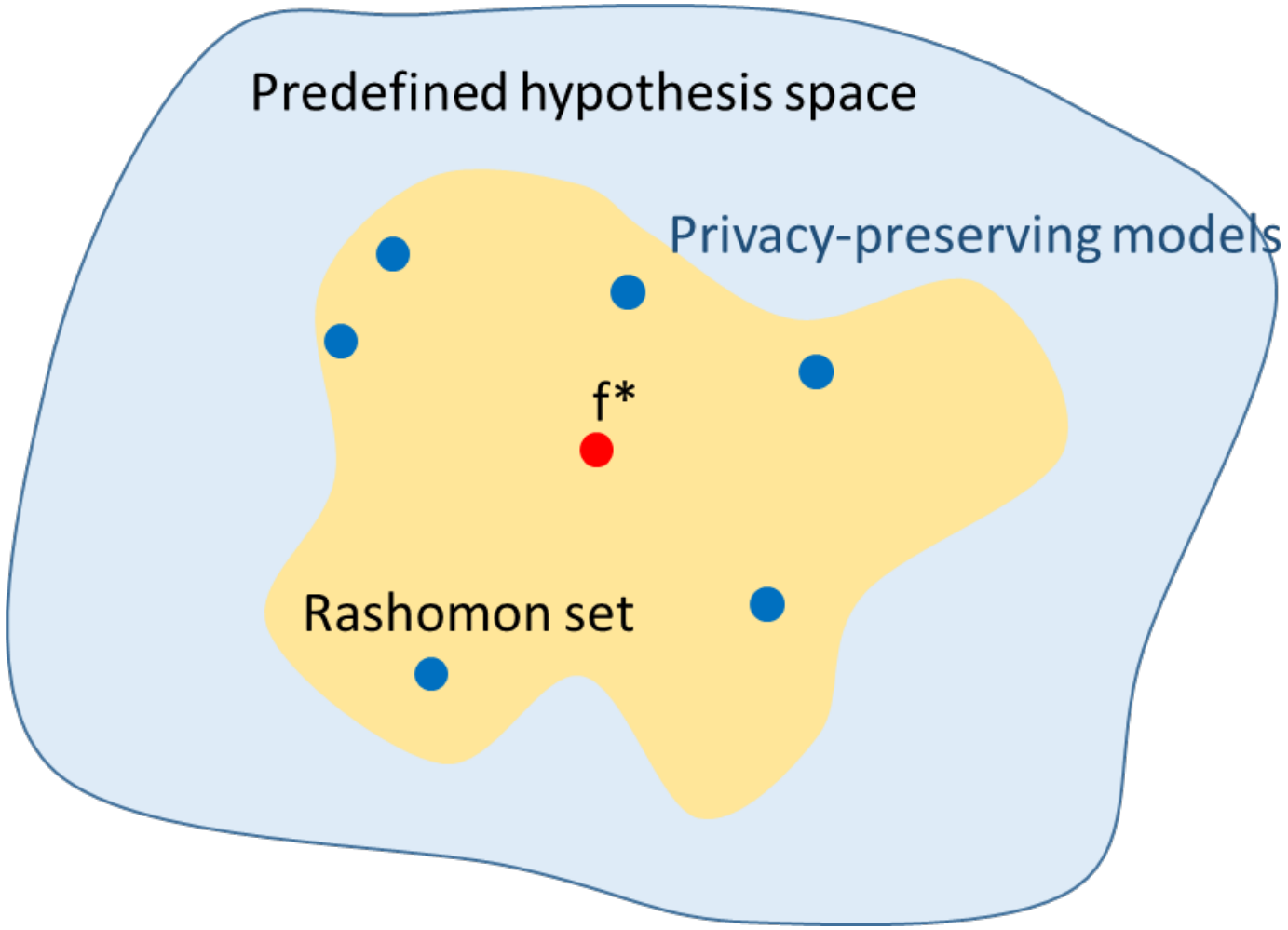
Slide by Chudi Zhong



Slide by Chudi Zhong

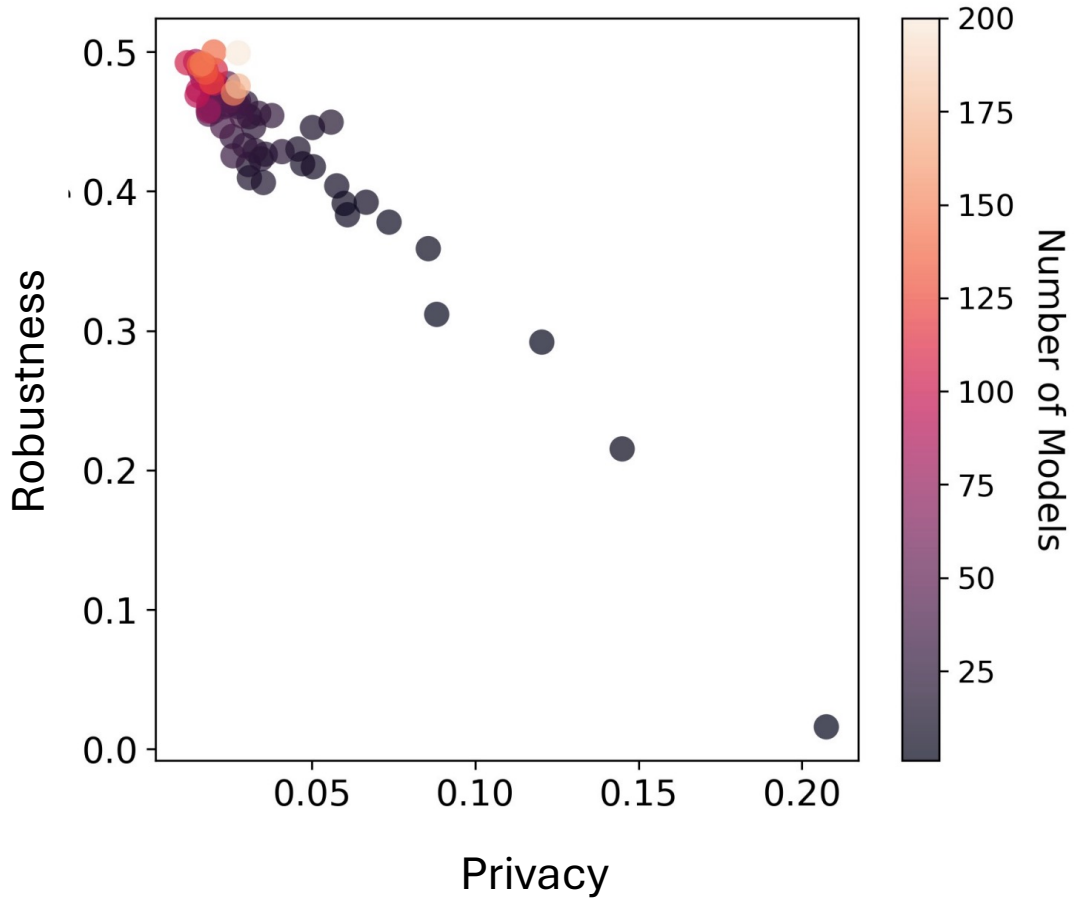


Slide by Chudi Zhong



Slide by Chudi Zhong

The Rashomon set itself should be considered as **policy object** as some properties can have trade-offs.



The Double-Edged Nature of the Rashomon Set for Trustworthy Machine Learning

ICML Spotlight 2026

Ethan Hsu[1*] Harry Chen[2*] Chudi Zhong[3] Lesia Semenova[4]
[1]Duke University [2]MIT [3]UNC-Chapel Hill [4]Rutgers University

TAKEAWAY 2:

In high-stakes decision domains

simpler (interpretable) models

should be the default.

Interpretability: Models that are inherently constrained so that their reasoning is understandable to humans

help to identify and mitigate the potential biases

The New York Times

Opinion

OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

June 13, 2017

Share full article 230

Sally Deng

are easier to troubleshoot

Innovate Healthcare

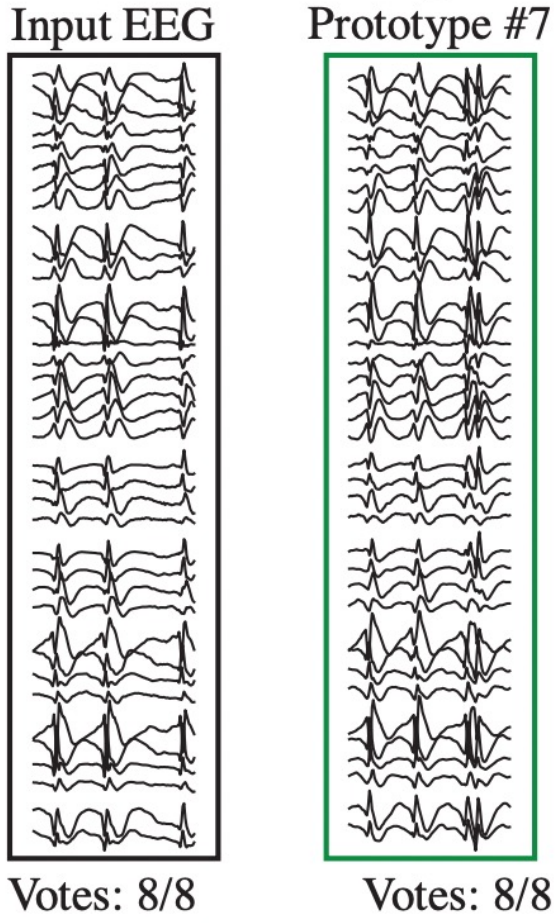
HealthImaging

INSIGHTS IN IMAGING & INFORMATICS

Algorithm's 'unexpected' weakness raises larger concerns about AI's potential in broader populations

Matt O'Connor | April 05, 2021 | Health Imaging | Artificial Intelligence

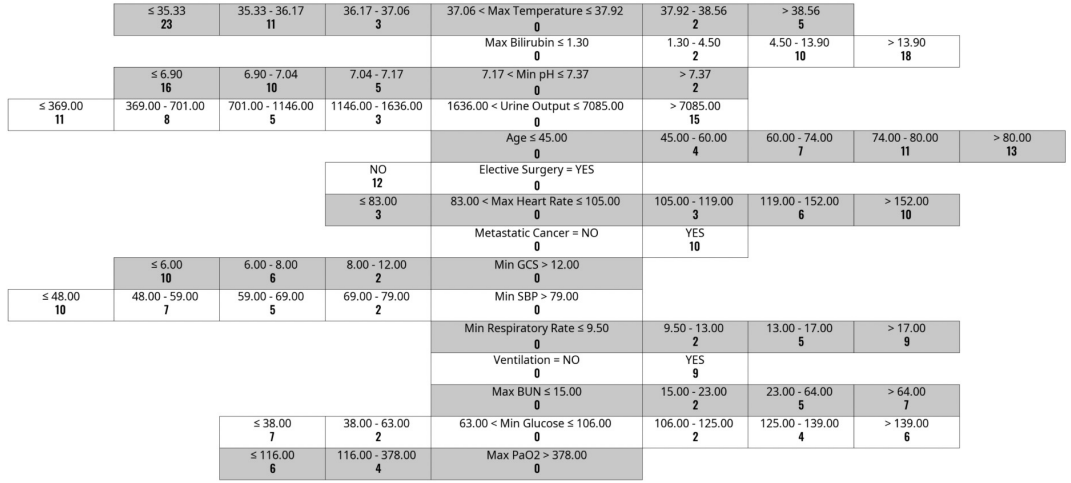
Seizure detection



Tang,.., Semenova, et al., NeurIPS workshop, 2023, oral

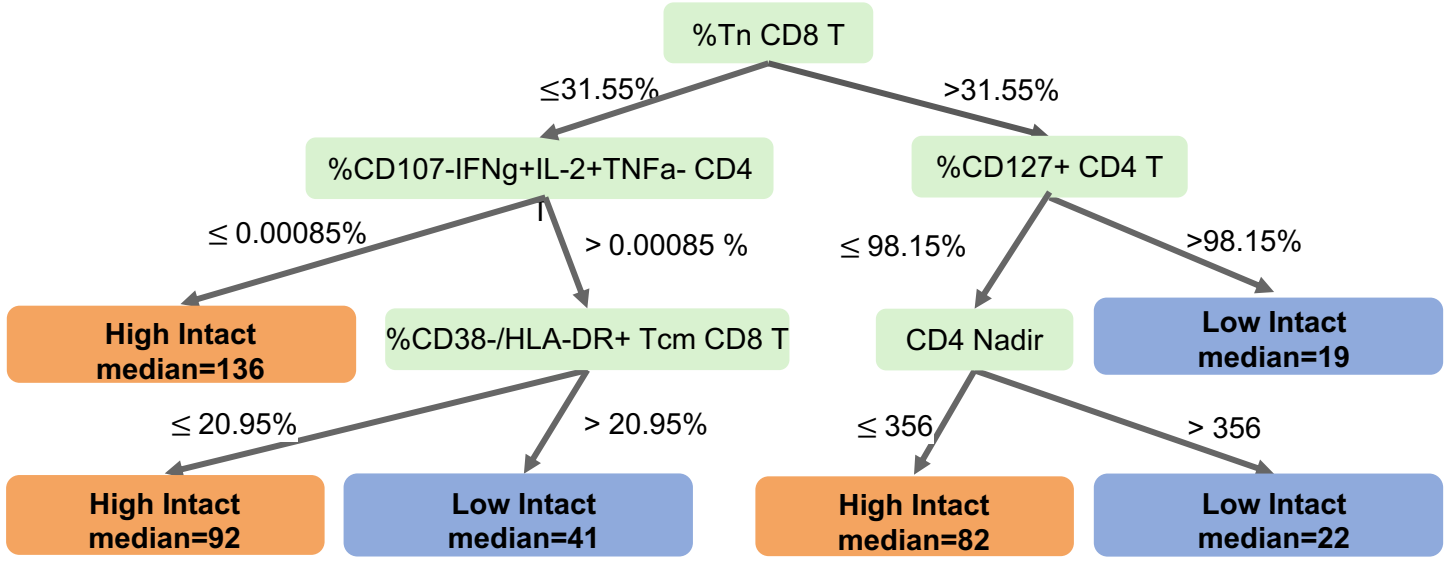
Mortality prediction

Score	4	8	11	15	18	22	25	29	32	36	39	43	46	50	53	56	60	63	67	70	74	77	81	84	88	91	95	98	102	109
Risk	0.1%	0.2%	0.2%	0.3%	0.5%	0.7%	1.1%	1.5%	2.1%	3.1%	4.7%	6.4%	8.6%	12.7%	18.2%	23.5%	29.8%	39.4%	50.0%	58.0%	65.6%	74.5%	81.8%	86.1%	89.5%	92.9%	95.3%	96.5%	97.5%	98.9%



Zhu, Tian, Semenova, et. al., 2023

HIV reservoir prediction



Semenova et. al., eLife, 2024

On the Existence of Simpler Machine Learning Models

Lesia Semenova, Cynthia Rudin, and Ronald Parr

{*lesia, cynthia, parr*}@cs.duke.edu

Department of Computer Science, Duke University

FAccT 2022

A Path to Simpler Models Starts With Noise

NeurIPS 2023

Lesia Semenova **Harry Chen** **Ronald Parr** **Cynthia Rudin**
Department of Computer Science, Duke University
{lesia.semenova,harry.chen084,ronald.parr,cynthia.rudin}@duke.edu

Regularized misclassification error with random label noise ρ

Given $\rho < 0.5$, each label is flipped with probability ρ, meaning

$$P(y' \neq y) = \rho,$$

where y' is a new label.

Theorem 8, Boner, Chen, Semenova, et al., NeurIPS 2024

Regularized misclassification error with random label noise ρ

$$\underset{f \in F}{\operatorname{argmin}}\ L\left(D_{\boldsymbol{noisy}}, f\right) + \boldsymbol{\lambda}\, R(f) = \underset{f \in F}{\operatorname{argmin}}\ L\left(D_{\boldsymbol{clean}}, f\right) + \frac{\boldsymbol{\lambda}}{\boldsymbol{1-2\rho}}\, R(f)$$

L – 0-1 loss, F – hypothesis space
R(f) – regularization function with regularization parameter λ
D – true distribution

Theorem 8, Boner, Chen, Semenova, et al., NeurIPS 2024

Regularized misclassification error with random label noise ρ

$$\underset{f \in F}{\operatorname{argmin}}\ L\left(D_{\boldsymbol{noisy}}, f\right) + \boldsymbol{\lambda}\, R(f) = \underset{f \in F}{\operatorname{argmin}}\ L\left(D_{\boldsymbol{clean}}, f\right) + \frac{\boldsymbol{\lambda}}{\boldsymbol{1-2\rho}} R(f)$$

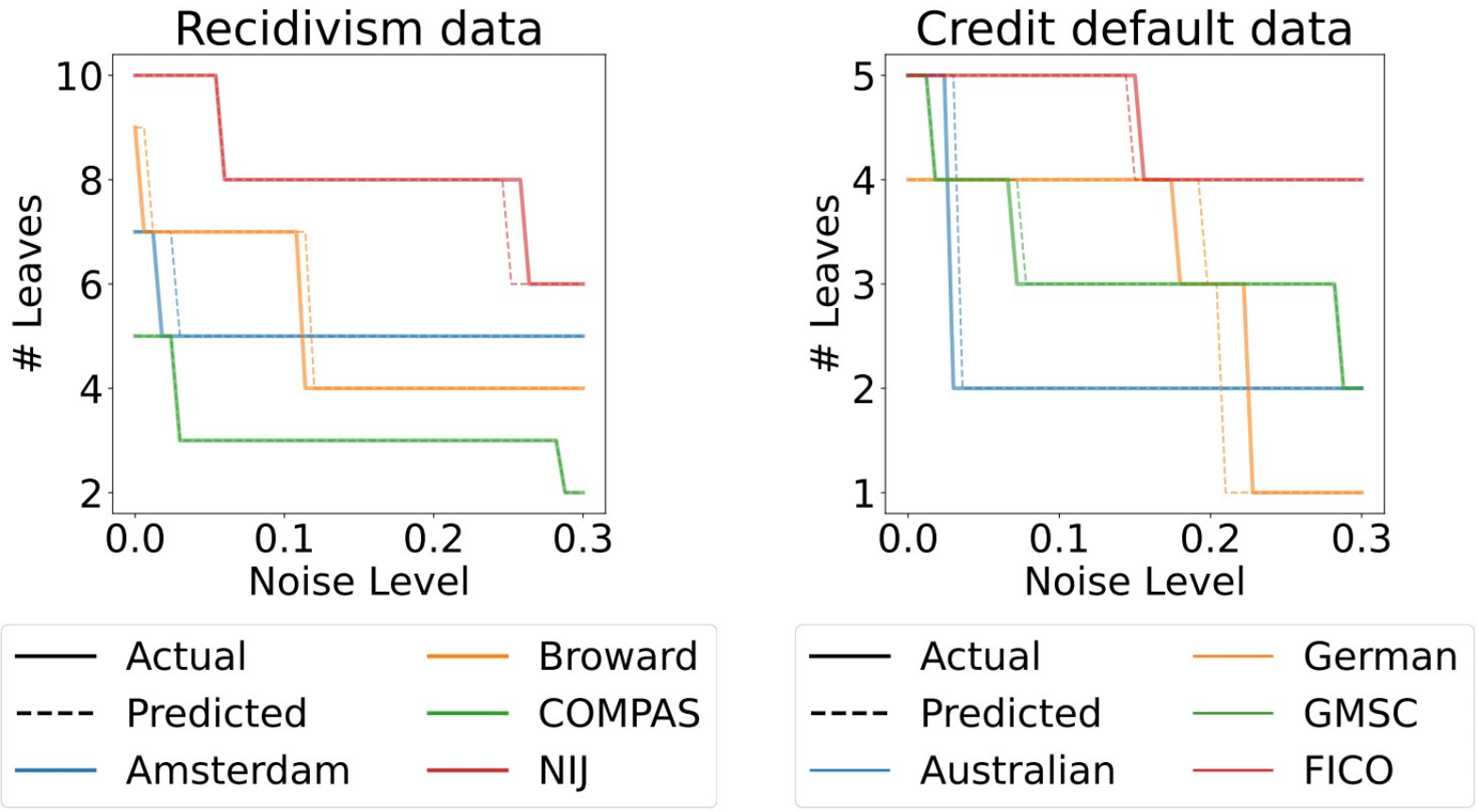
L – 0-1 loss, F – hypothesis space
R(f) – regularization function with regularization parameter λ
D – true distribution

Optimizing over the noisy data distribution is equivalent to

optimizing over the clean data distribution with **stronger regularization**

penalty $\lambda/1-2\rho$.

Theorem 8, Boner, Chen, Semenova, et al., NeurIPS 2024



Theorem 8, Boner, Chen, Semenova, et al., NeurIPS 2024

TAKEAWAY 3

We can use **knowledge of the Rashomon set**

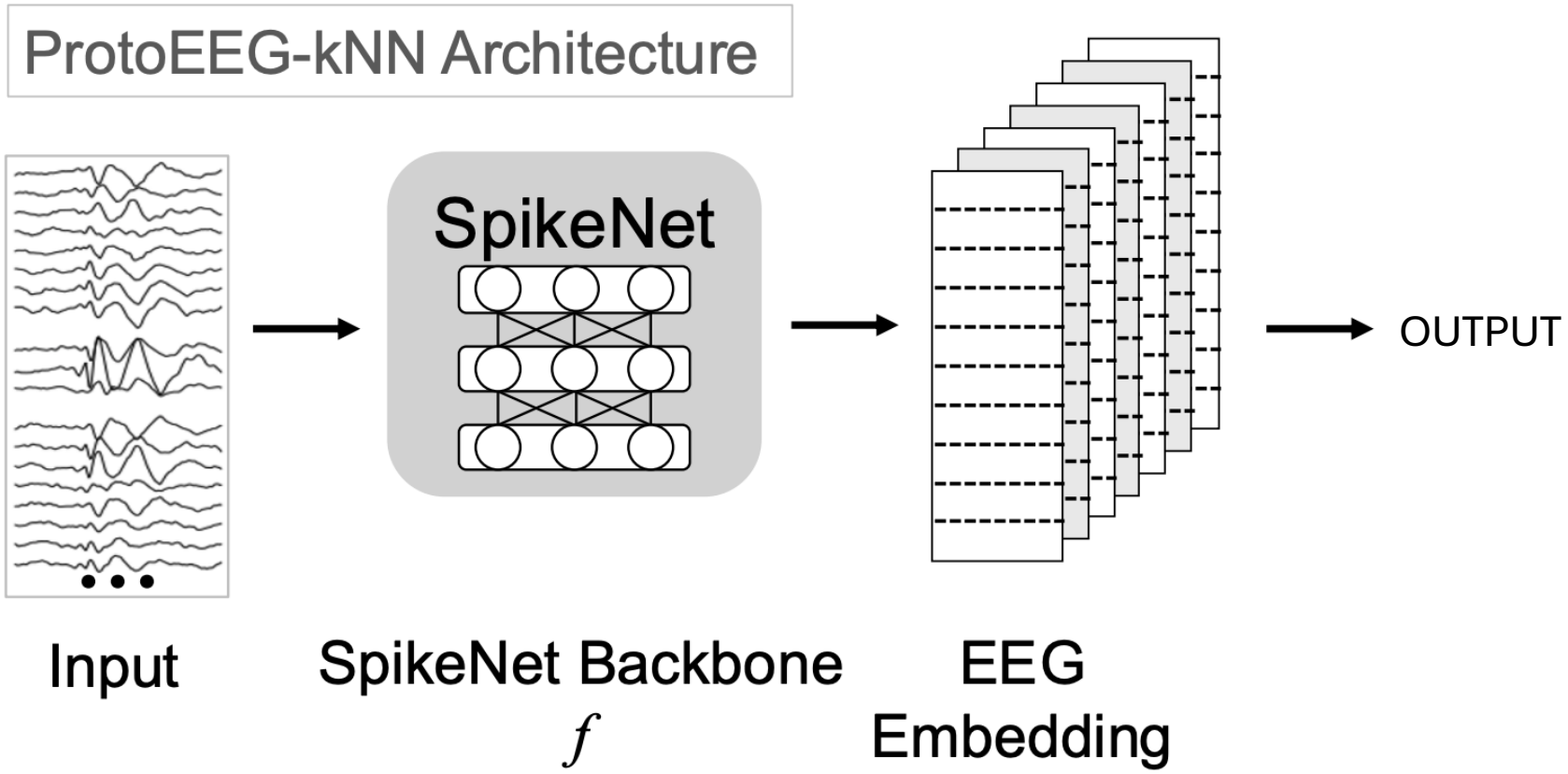
to design better algorithms for trustworthy AI,

even if the set was not computed directly

Explainability - post-hoc interpretation of a model’s decision

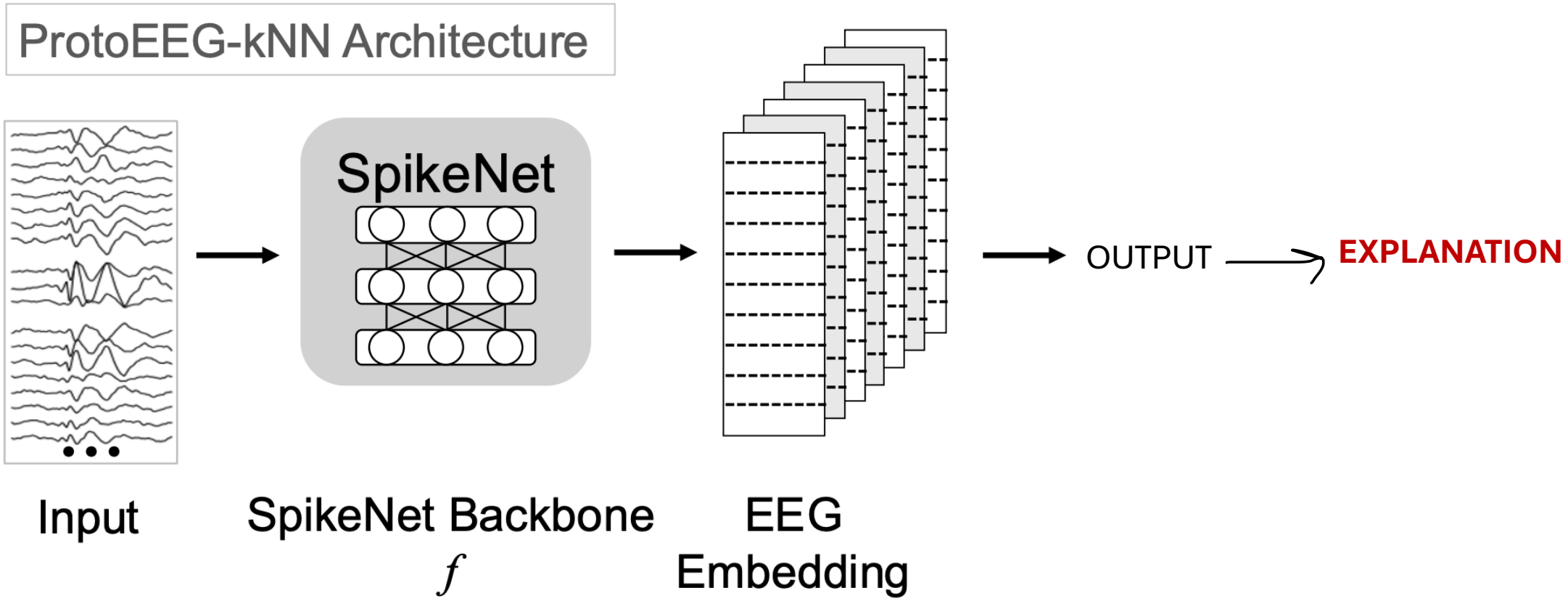
Explainability - post-hoc interpretation of a model's decision

Black box architecture for spike detection

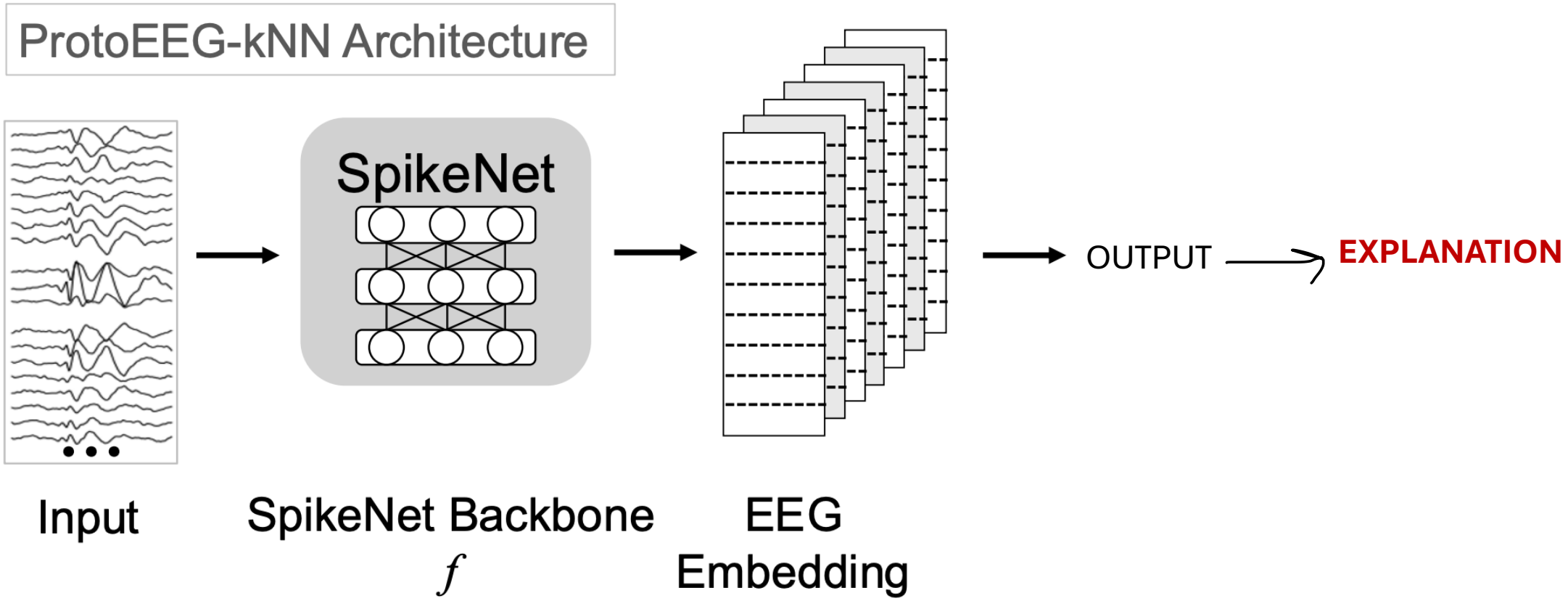


Explainability - post-hoc interpretation of a model's decision

Black box architecture for spike detection

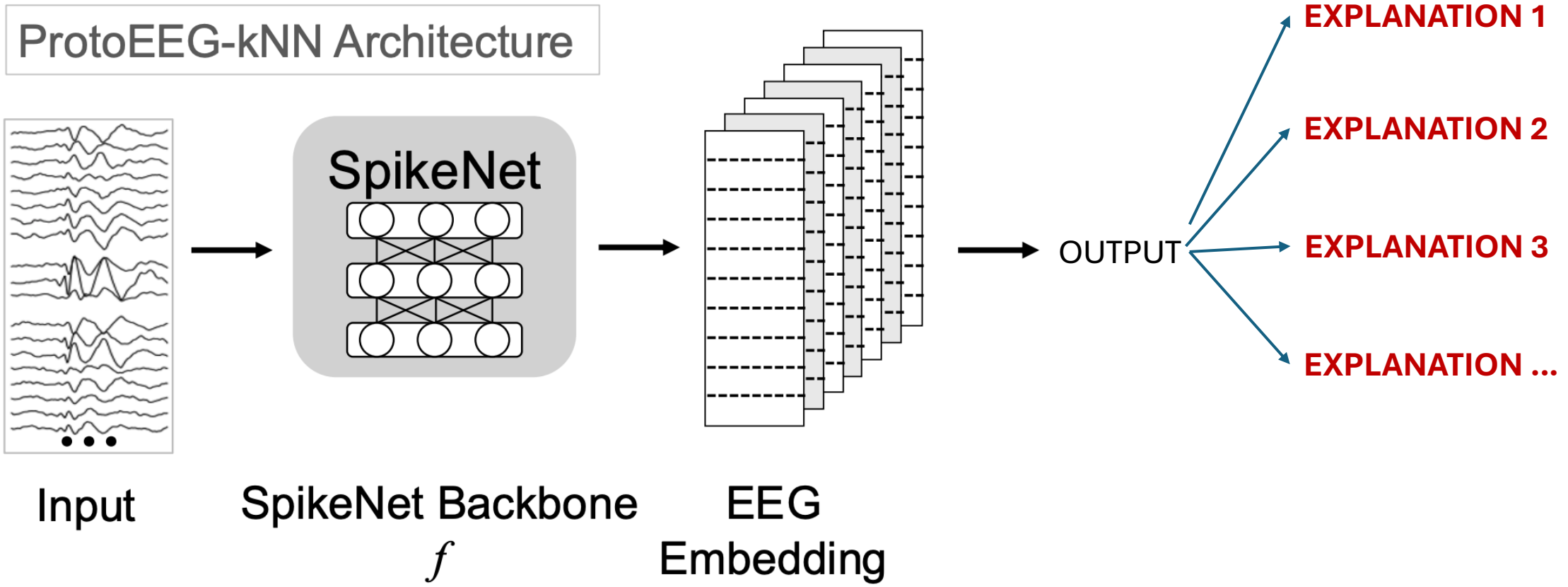


Explainability - post-hoc interpretation of a model's decision



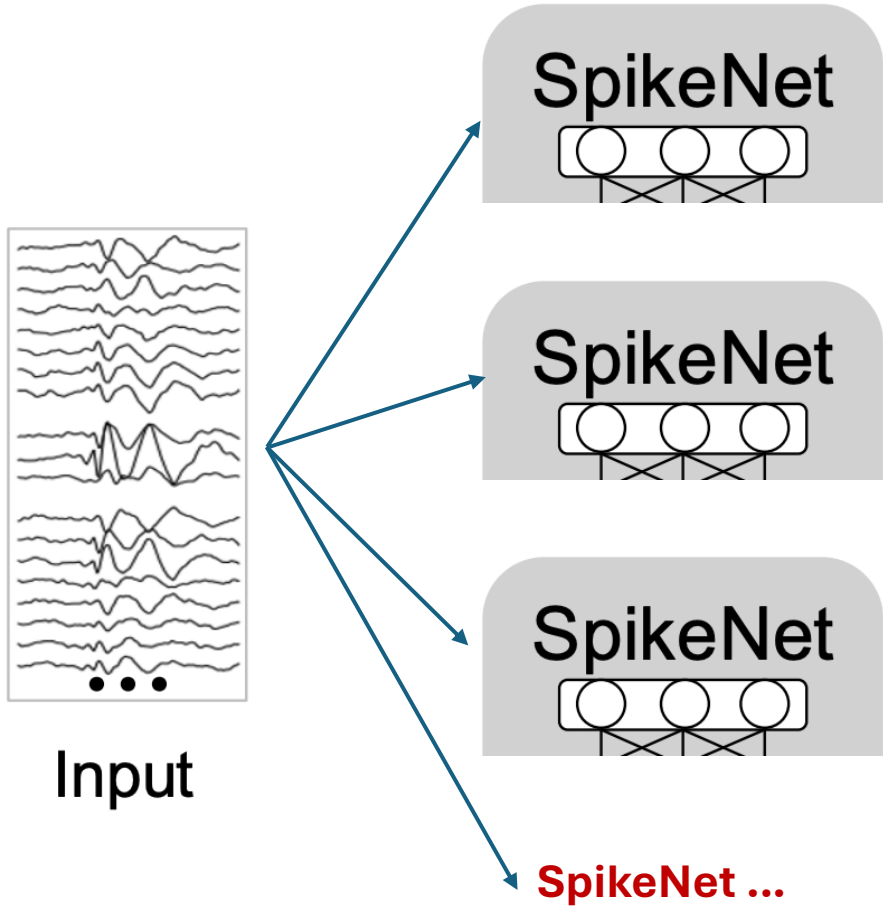
Explainability - post-hoc interpretation of a model's decision

Rashomon Effect occurs at least two levels: multiplicity in **explanations**.



Explainability - post-hoc interpretation of a model's decision

Rashomon Effect occurs at least two levels: multiplicity in **models**.



Explainability - post-hoc interpretation of a model's decision

- Multiplicity in explanations allows to choose the interpretation that **aligns with the users/stakeholders** (Pawelczyk et al., UAI 2020; Mothilal et al., FAccT 2020).

- Model multiplicity needs to be taken into consideration to ensure that explanations **remain valid** under model change

We designed fast, efficient robust recourse algorithm that is **immune to the model changes** by approximating the Rashomon set with ellipsoid (we do not compute it directly).

ElliCE: Efficient and Provably Robust Algorithmic Recourse via the Rashomon Sets

NeurIPS Spotlight 2025

Bohdan Turbal[1] **Iryna Voitsitska**[2] **Lesia Semenova**[3*]

TAKEAWAYS: **Source of uncertainty -> foundation for trust**

1. If we have large Rashomon set, we can simply **search it for the desirable property**

2. In high-stakes decision domains **simpler (interpretable) models** should be the **default**.

3. We can use **knowledge of the Rashomon set** to design better algorithms for trustworthy AI

Future directions and open problems

- What causes/determines Rashomon-set size?
- How do we find/represent Rashomon sets at scale?
  - Especially for more complex model classes such as LLMs or VLMs
- How to display the Rashomon set for human-model interaction?
- How to address individual arbitrariness in deployment?
- How does uncertainty (from data or modeling) influences AI systems?

… and many more

Keep building Rashomon set aware pipelines and algorithms!

Thank you!