To Stay or Not to Stay in the Pre-train Basin: Insights on Ensembling in Transfer Learning

Ildus Sadrtdinov*

Dmitrii Pozdeev*

Dmitry Vetrov

Ekaterina Lobacheva

HSE University

Technical University of Munich TUM

CONSTRUCTOR UNIVERSITY

p(B|A)yesgroup

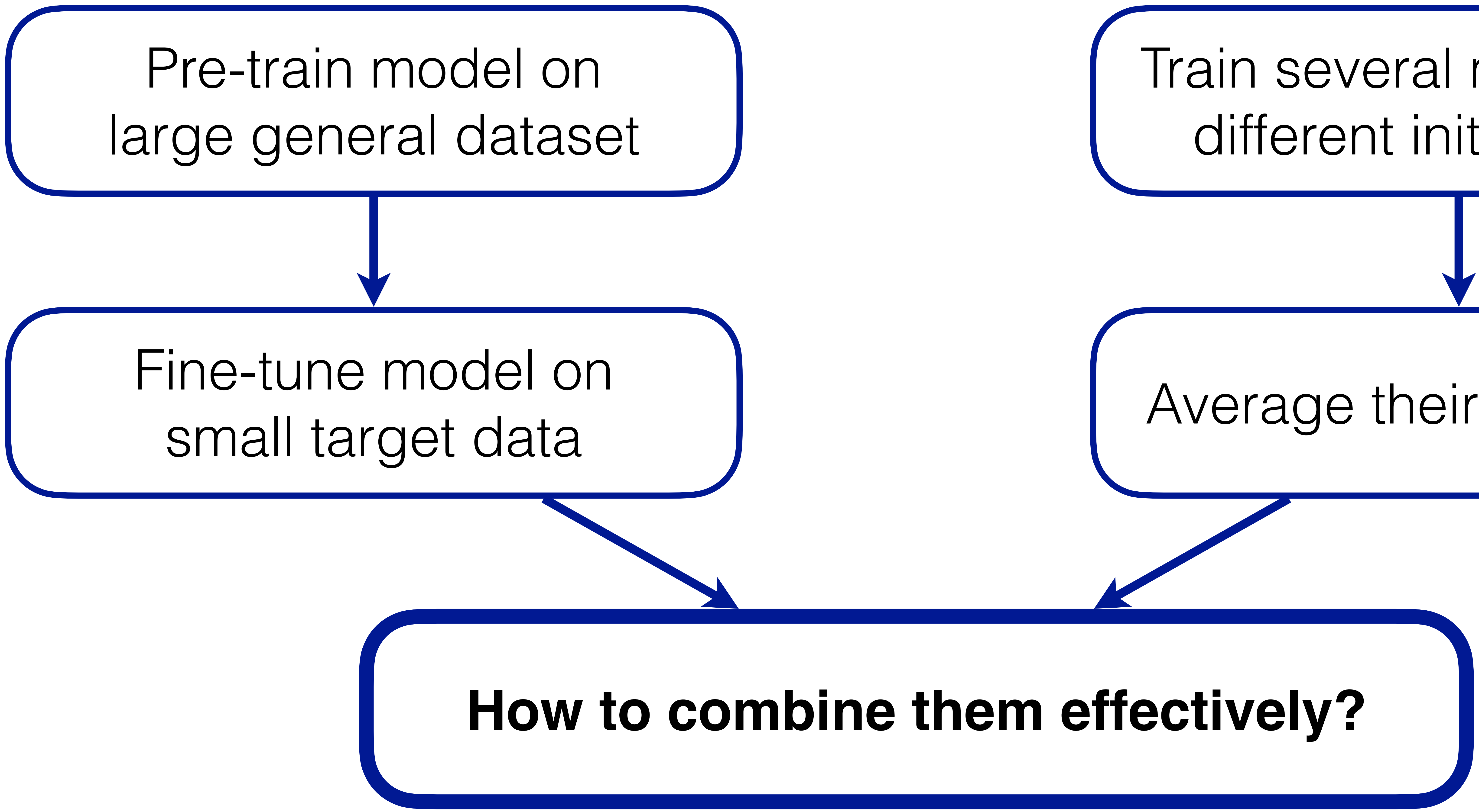
NEURAL INFORMATION PROCESSING SYSTEMS

* equal contribution

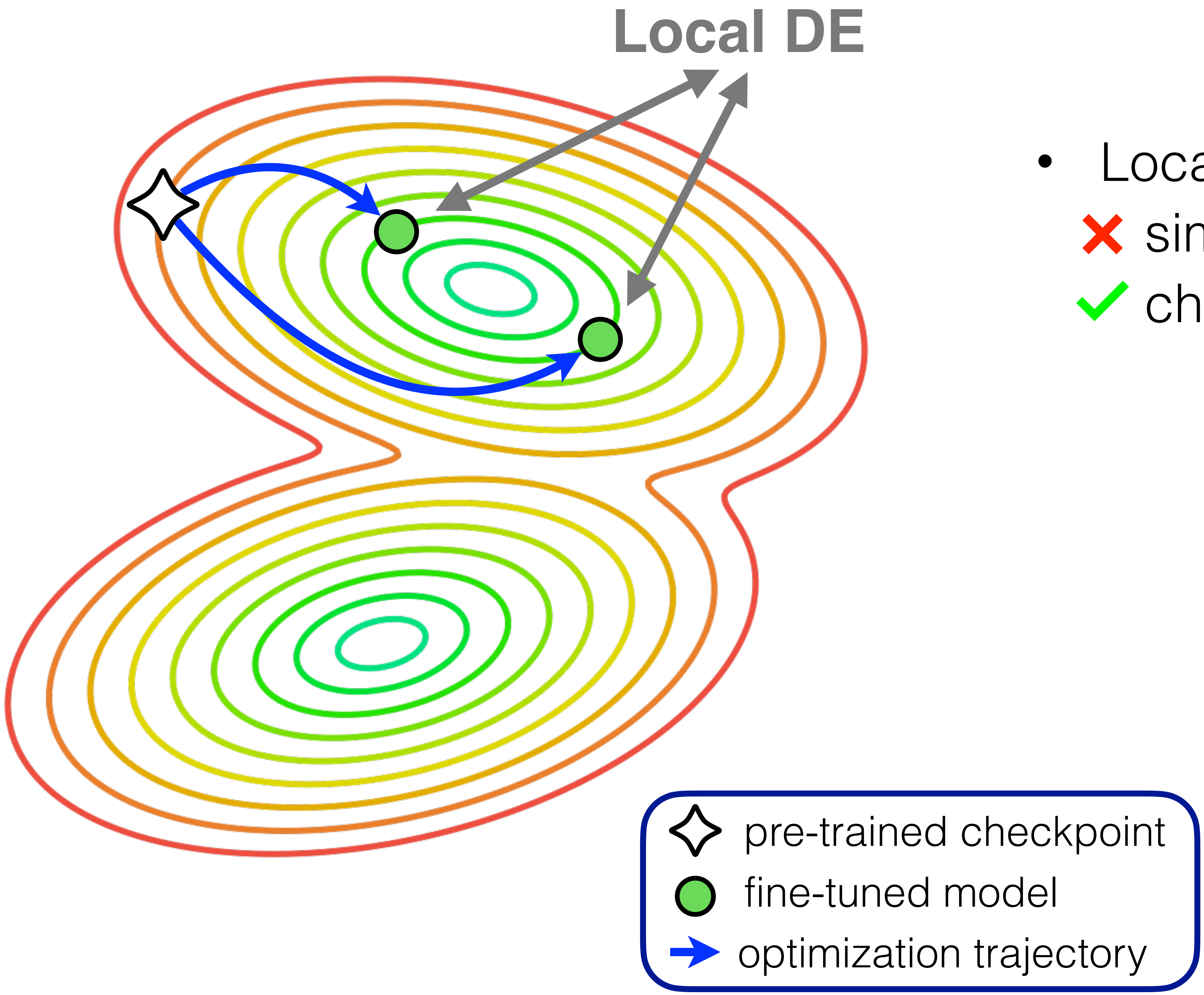
Transfer learning & ensembles

Transfer learning

Deep ensembles

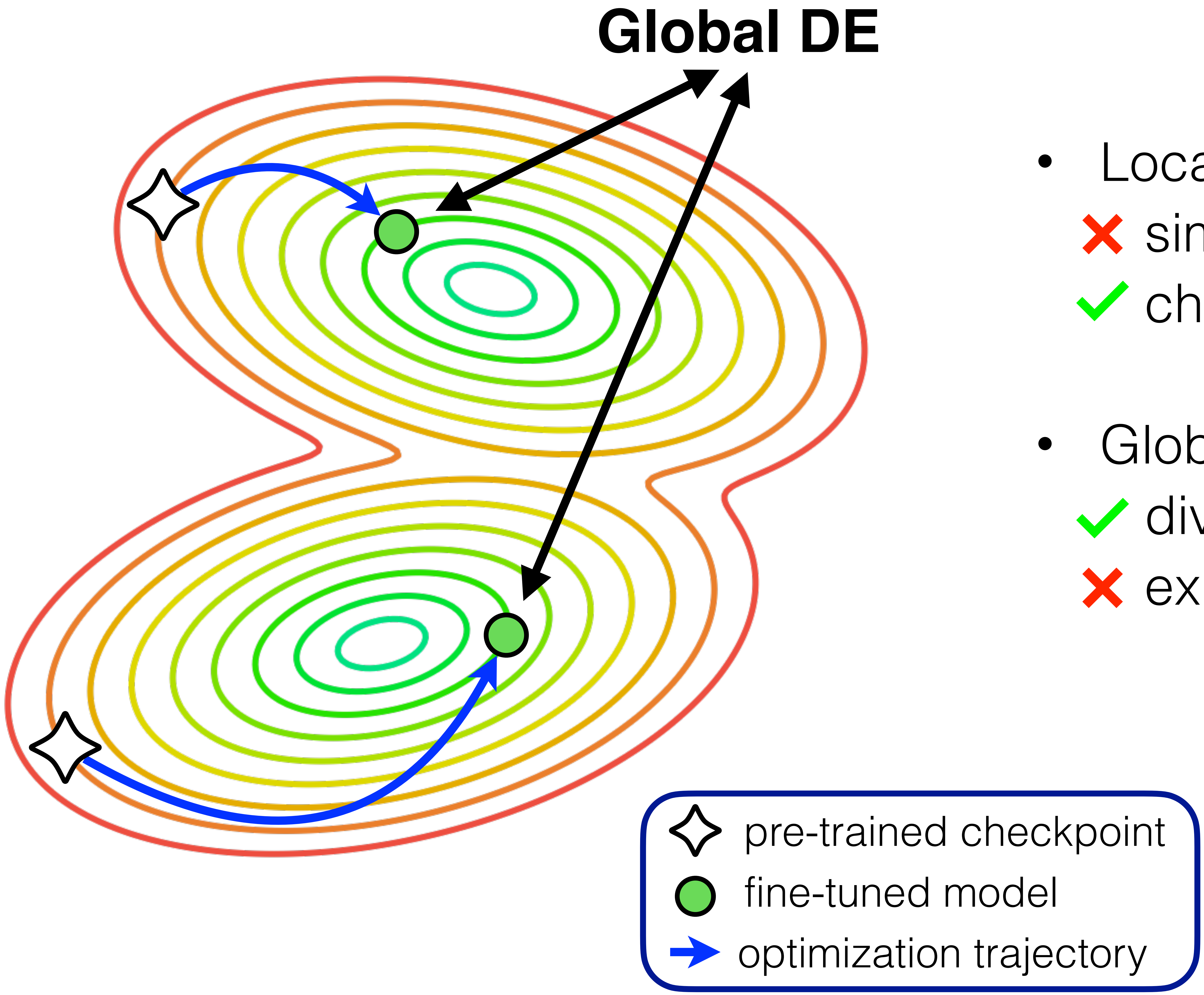


Ensembles in transfer learning



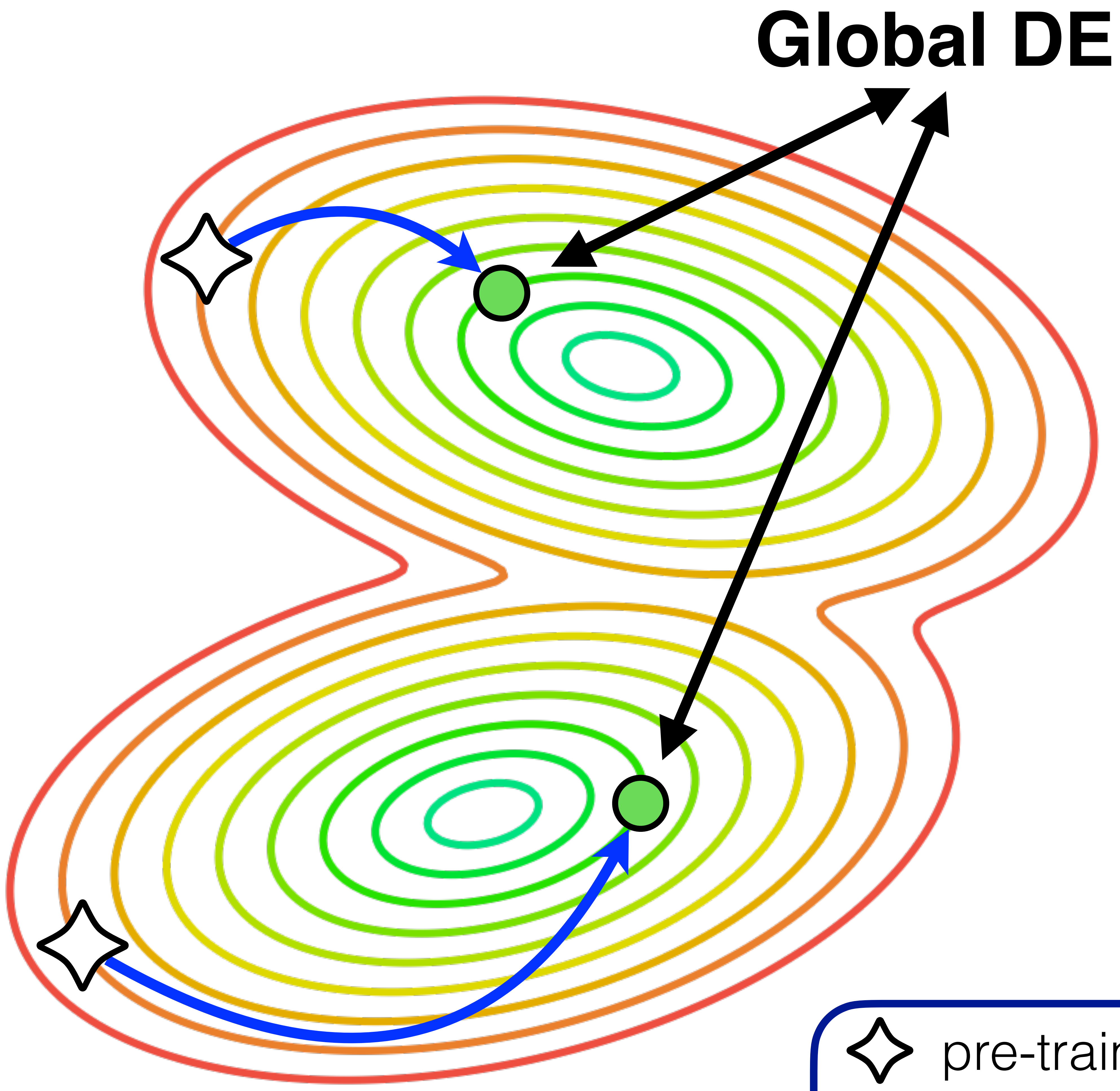
- Local Deep Ensemble (**Local DE**)
  ❌ similar networks, lower quality
  ✅ cheap to train

Ensembles in transfer learning

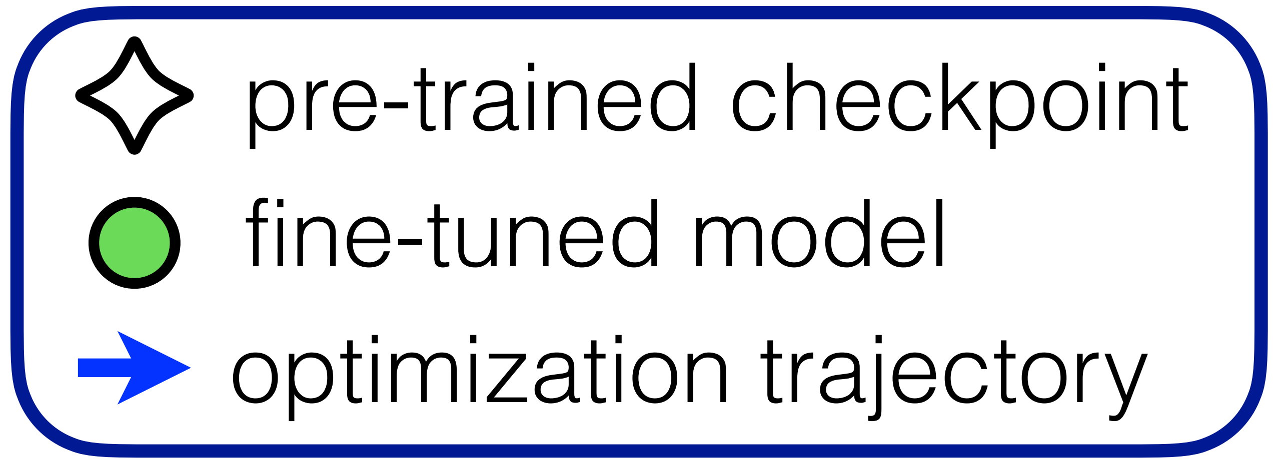


- Local Deep Ensemble (**Local DE**)
  - ✗ similar networks, lower quality
  - ✓ cheap to train
- Global Deep Ensemble (**Global DE**)
  - ✓ diverse networks, higher quality
  - ✗ expensive to train

Ensembles in transfer learning



Reduce the gap between **Local** and **Global DE** with one pre-trained model?



Experimental setup

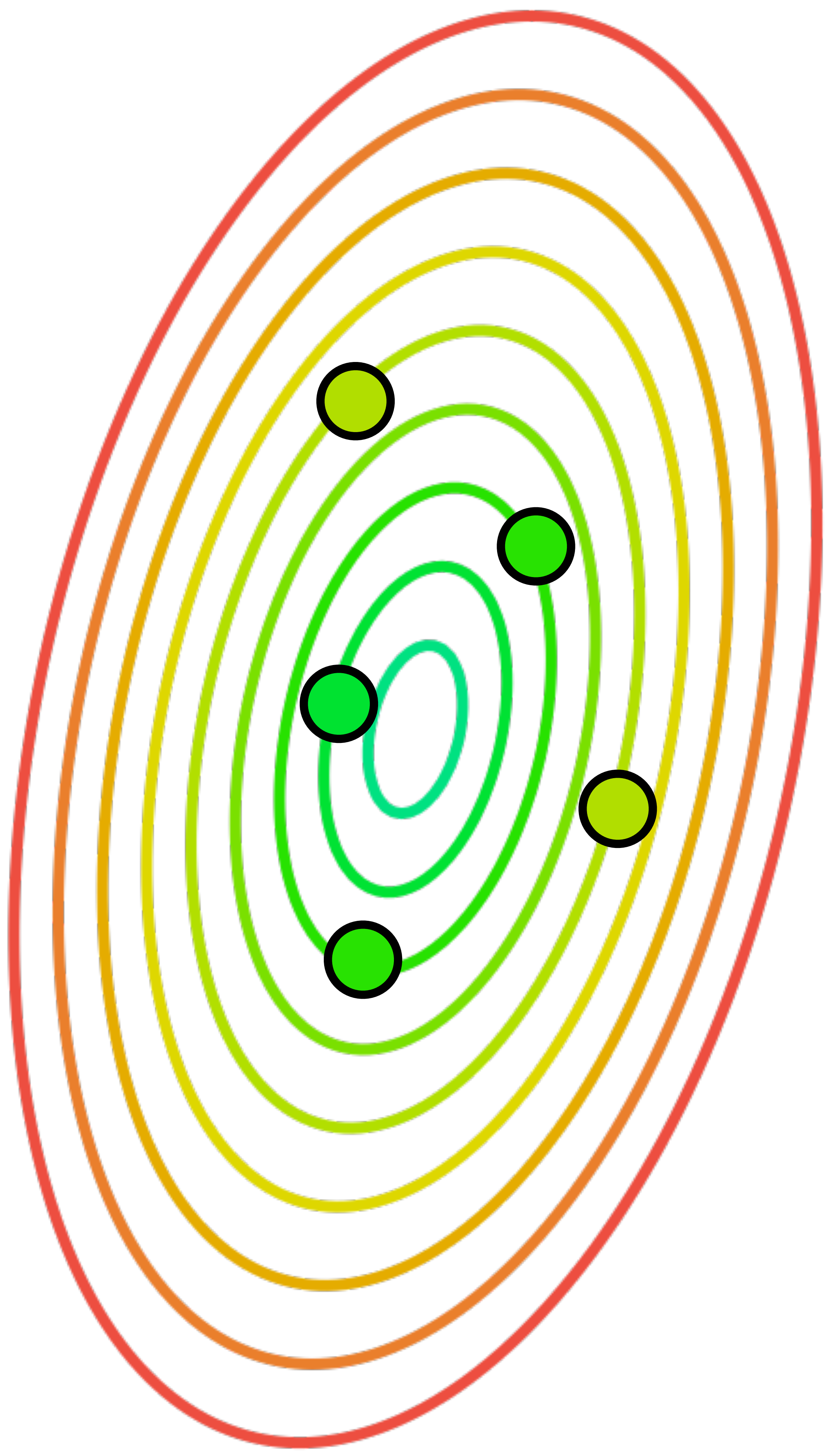
Architecture / pre-training type:

- ResNet-50 / BYOL (Grill et al, 2020) on ImageNet
- ResNet-50 / supervised on ImageNet
- Swin-T / supervised on ImageNet
- ViT-B/32 / CLIP

Fine-tuning datasets:

- Natural: CIFAR-10/100, SUN-397
- Non-natural: Chest-X, Clipart
- ImageNet (for CLIP pre-training only)

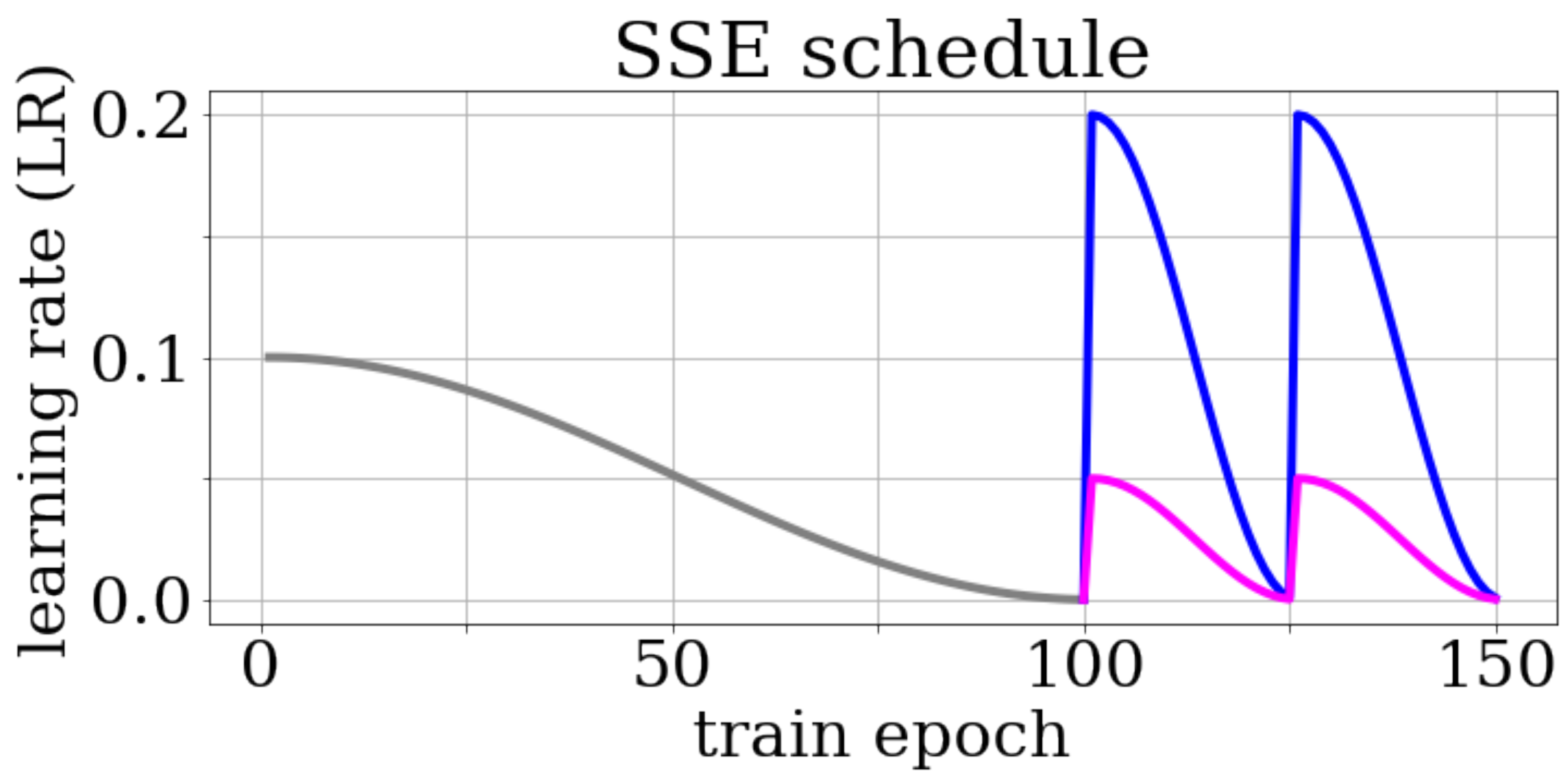
Effective ensembles in non-transfer setup



Possible approaches:

- Approximate the basin with some distribution and sample from it:
  - KFAC Laplace
  - SWA-Gaussian
  - SPRO (simplexes)
- Explore the basin using cyclical LR:
  - FGE
  - SSE
  - cSGLD

Can existing methods help?



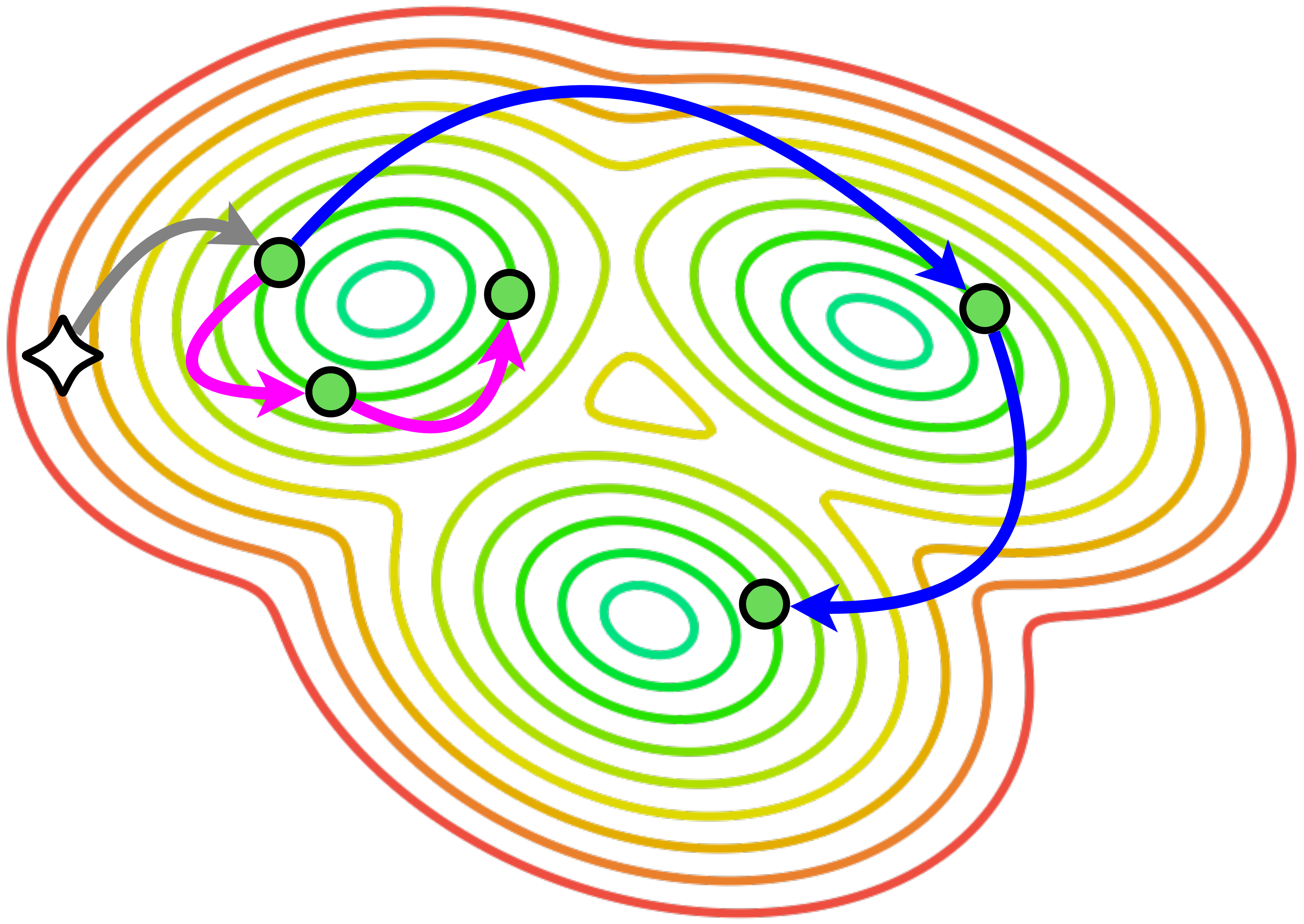
Cyclical methods, e.g. SnapShot Ensembles (SSE, Huang et al., 2017)

- Cyclical learning rate (LR) schedule, ensemble checkpoint at LR minima

Our experiments:

- **First network** — same as in Local DE
- **Following cycles** — different cycle hyperparametes (num epochs & max LR)

Local and semi-local behavior of SSE



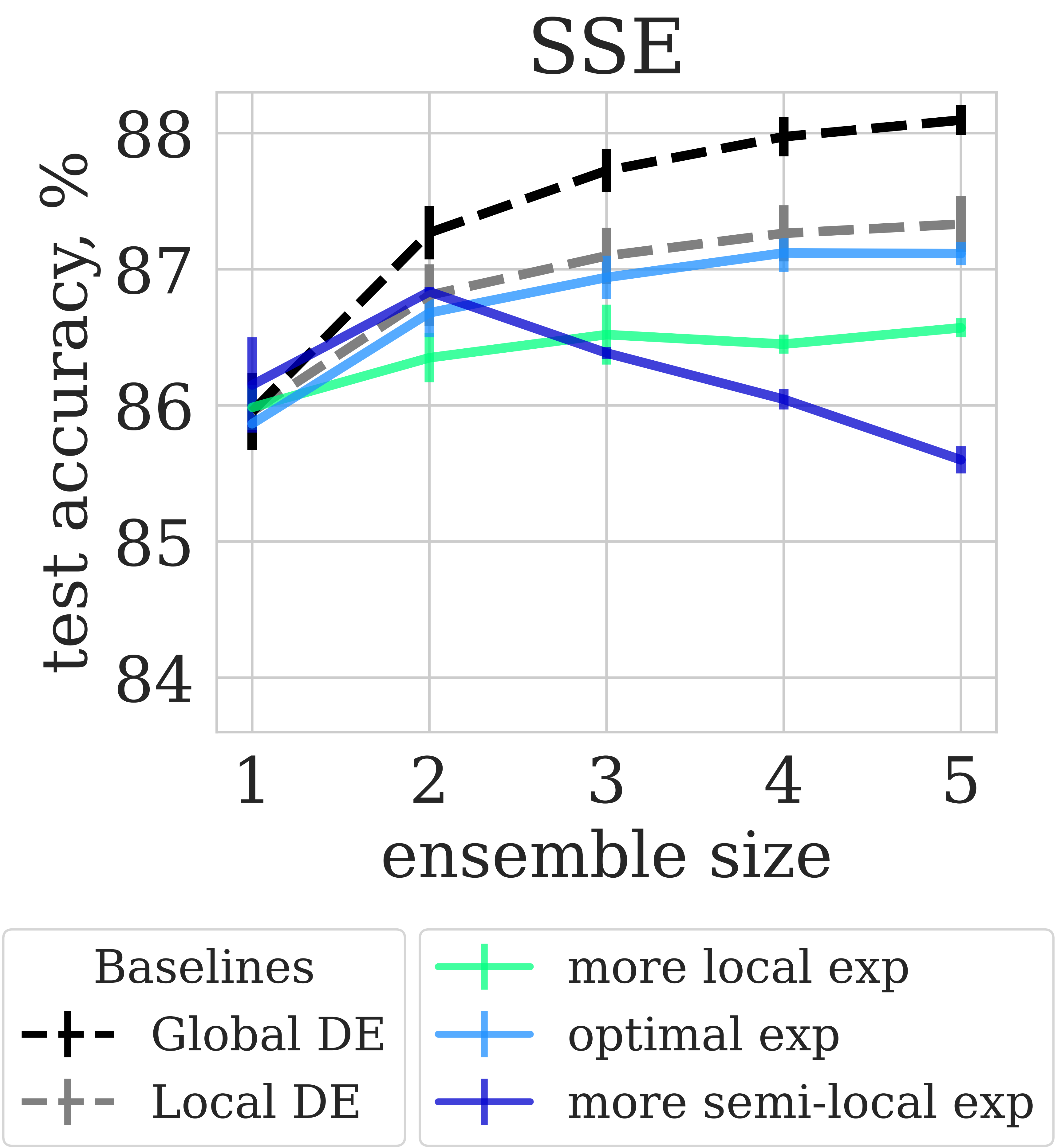
- Low hyperparameters → same basin → **local** behavior

- High hyperparameters → neighboring basins → **semi-local** behavior

Is it better to use SSE in a local or semi-local regime?

Finally, end of the problem setup…
Questions?

SSE results

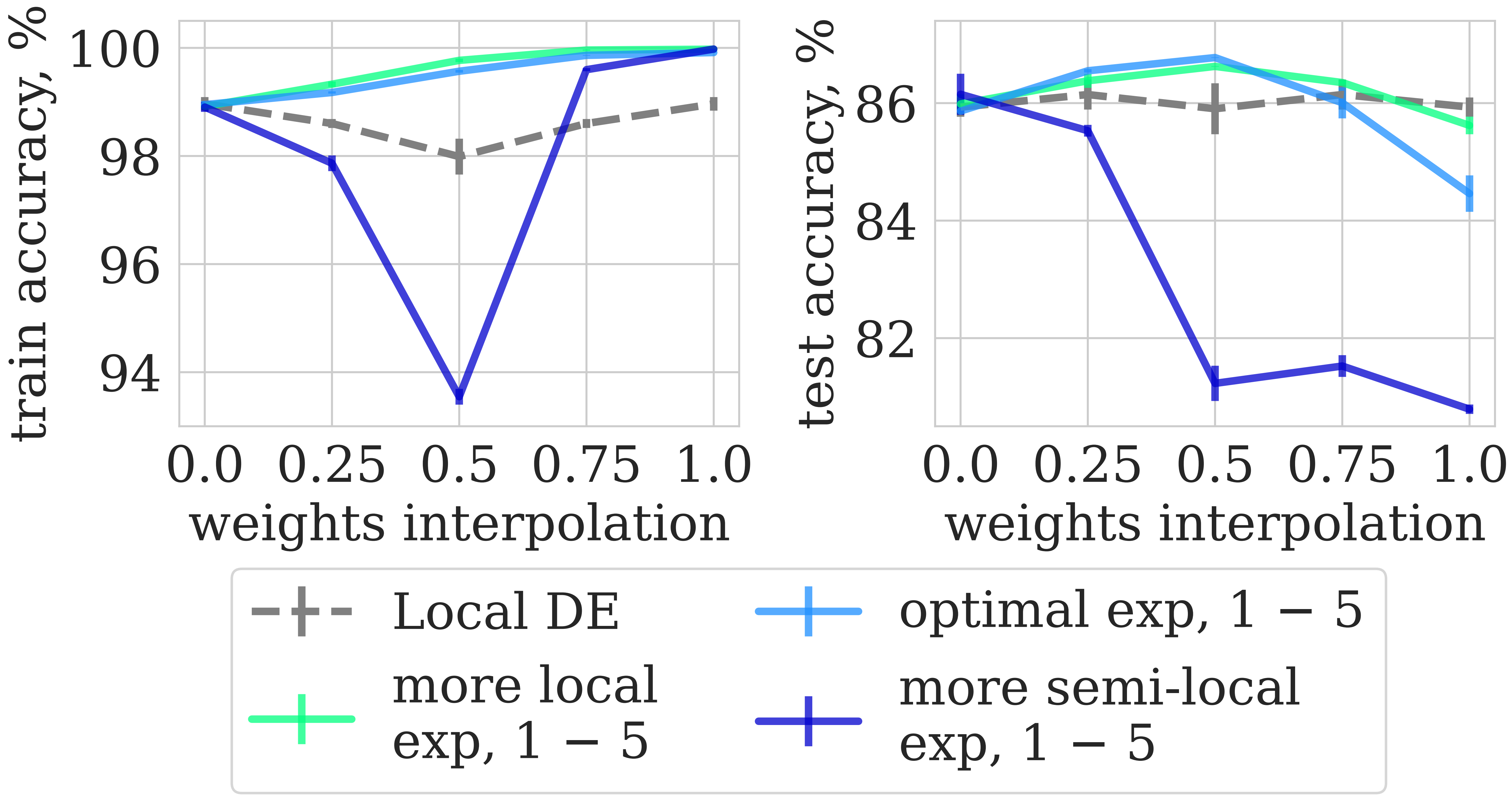


ResNet-50, CIFAR-100,
BYOL self-supervised pre-training.

3 main SSE results:

- **More local SSE** — models are very close, slowly growing ensemble quality
- **Optimal SSE** — more diverse models, quality comparable to Local DE
- **More semi-local SSE** — low quality of ensembles of larger sizes

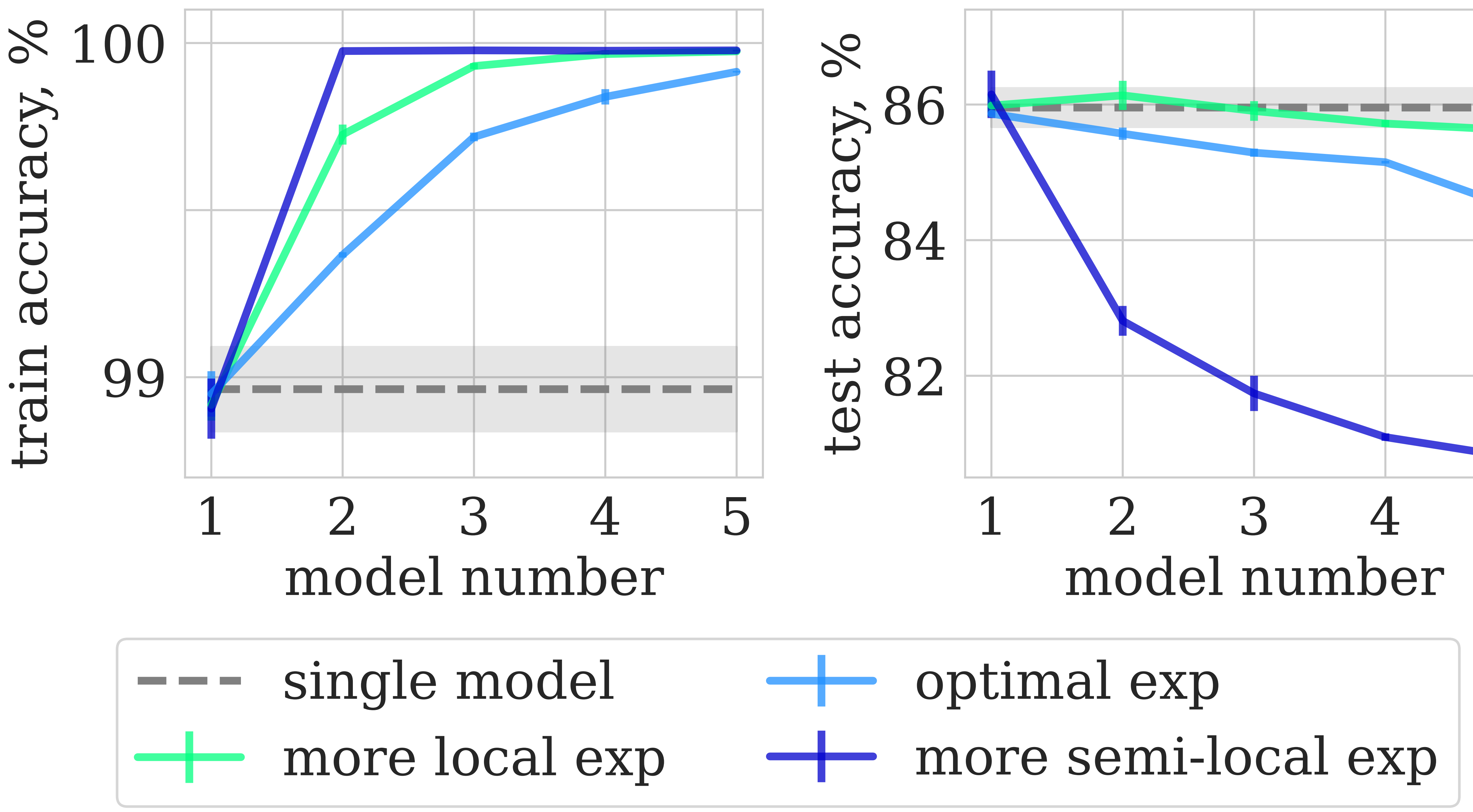
SSE analysis



ResNet-50, CIFAR-100,
BYOL self-supervised pre-training.

- **More local SSE** & **Optimal SSE** — local behavior (no accuracy drop in the middle, same basin)
- **More semi-local SSE** — semi-local behaviour (accuracy drop in the middle, different basins)

SSE analysis



ResNet-50, CIFAR-100,
BYOL self-supervised pre-training.

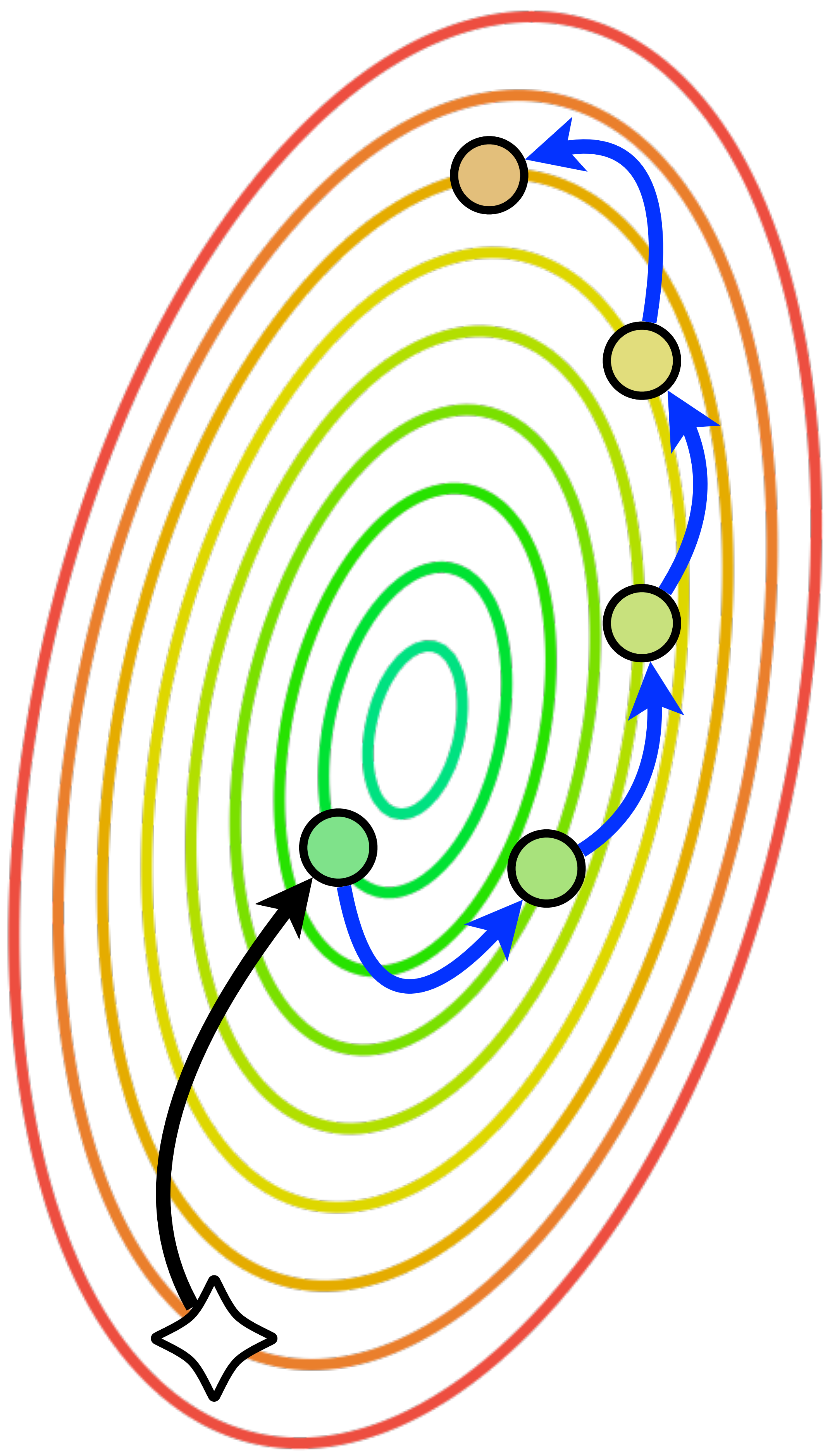
After each cycle:

- Train accuracy ↑
- Test accuracy ↓

SSE:

- overfits
- goes too far from pre-trained checkpoint
- loses advantages of transfer learning

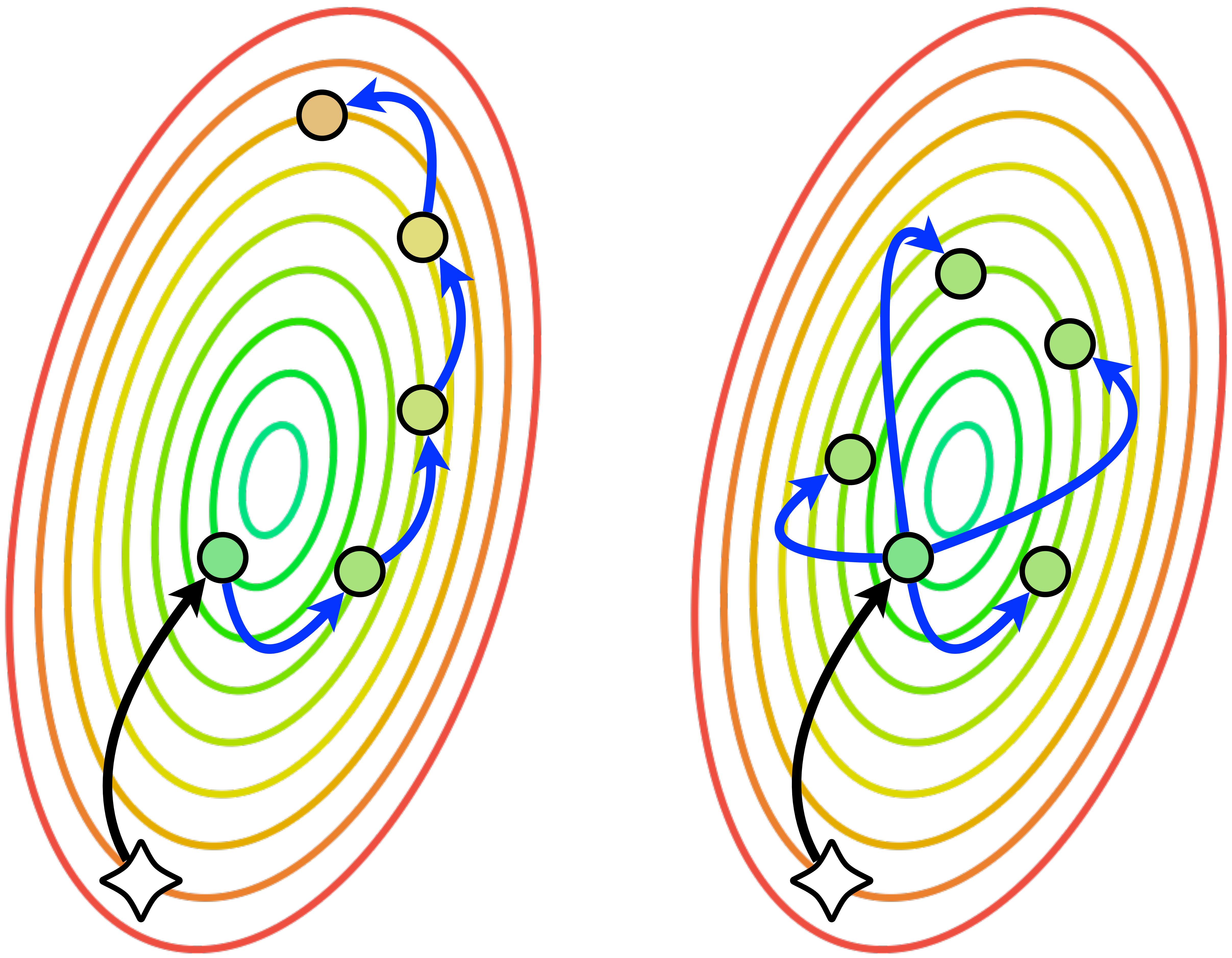
Can we do better than SSE?



SSE

- **Problem:** sequential training → degradation of models quality

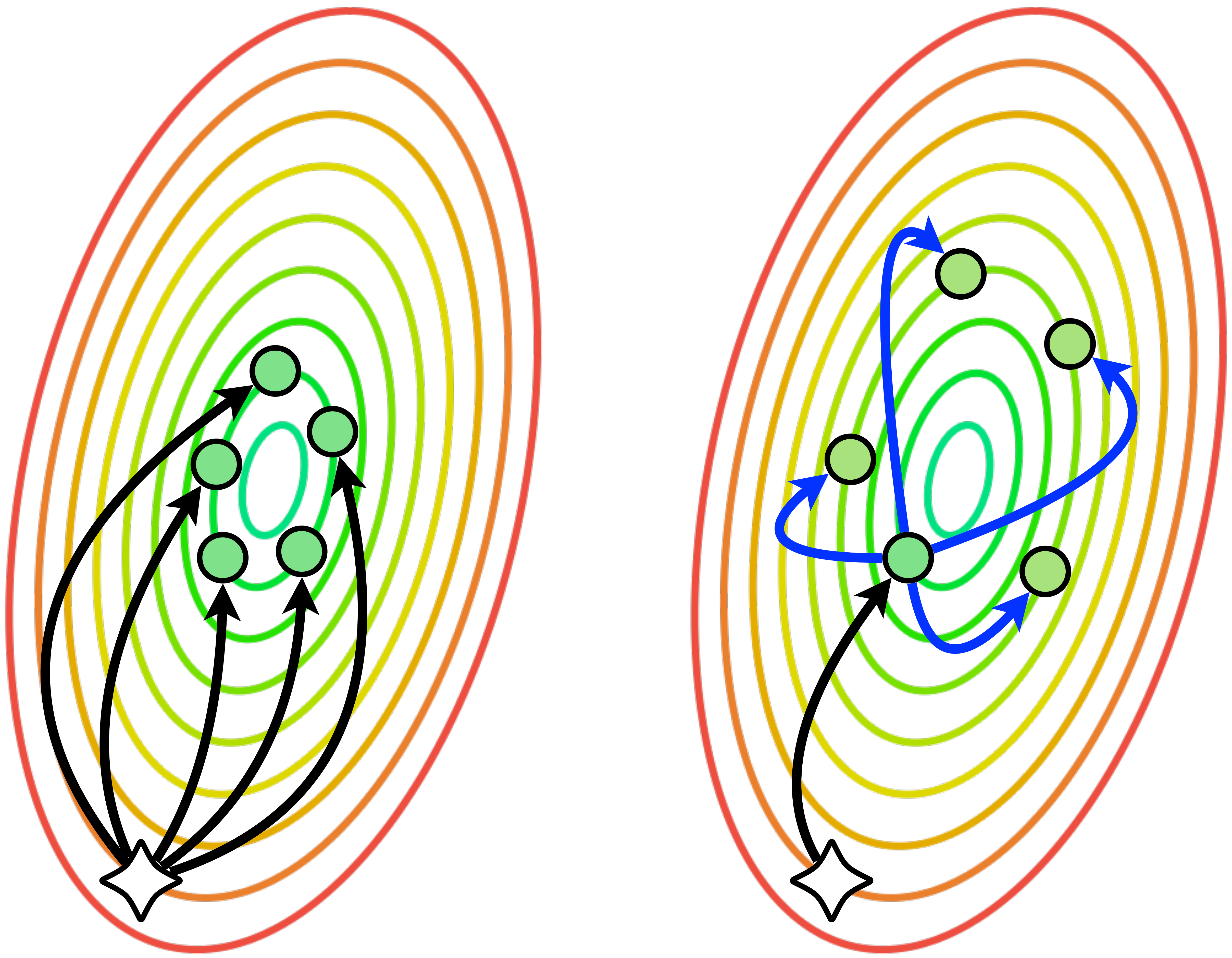
StarSSE, our modification of SSE



SSE StarSSE

- **Problem:** sequential training → degradation of models quality
- **Solution:** train models in parallel!
- First network trained similarly to SSE
- Rest of models trained in parallel starting from the first network

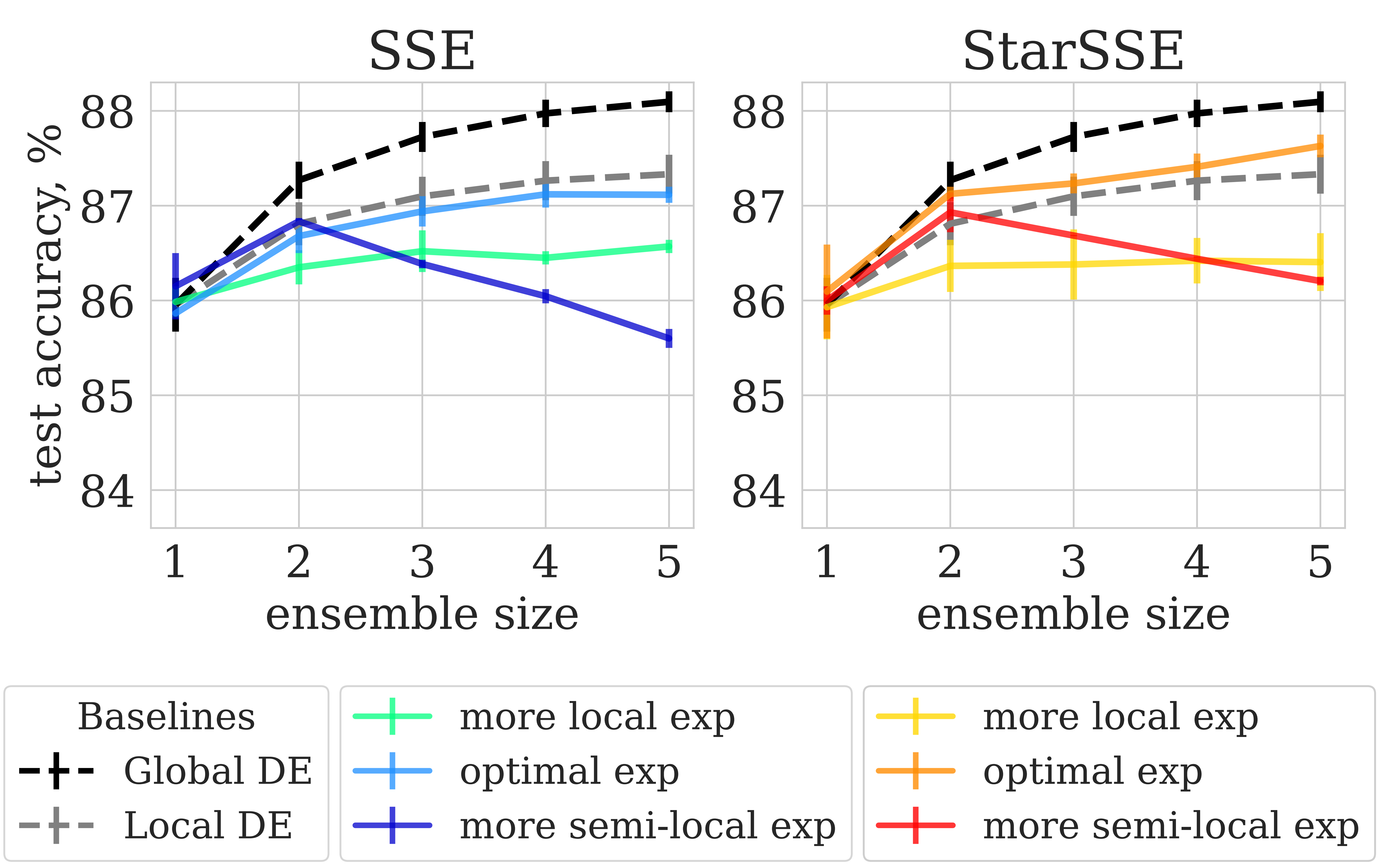
StarSSE and Local DE



Local DE StarSSE

- **Local DE:** parallel training from pre-trained checkpoint
- **StarSSE:** parallel training from fine-tuned model
- StarSSE separates **moving to low-loss region** and **pre-train basin exploration**!

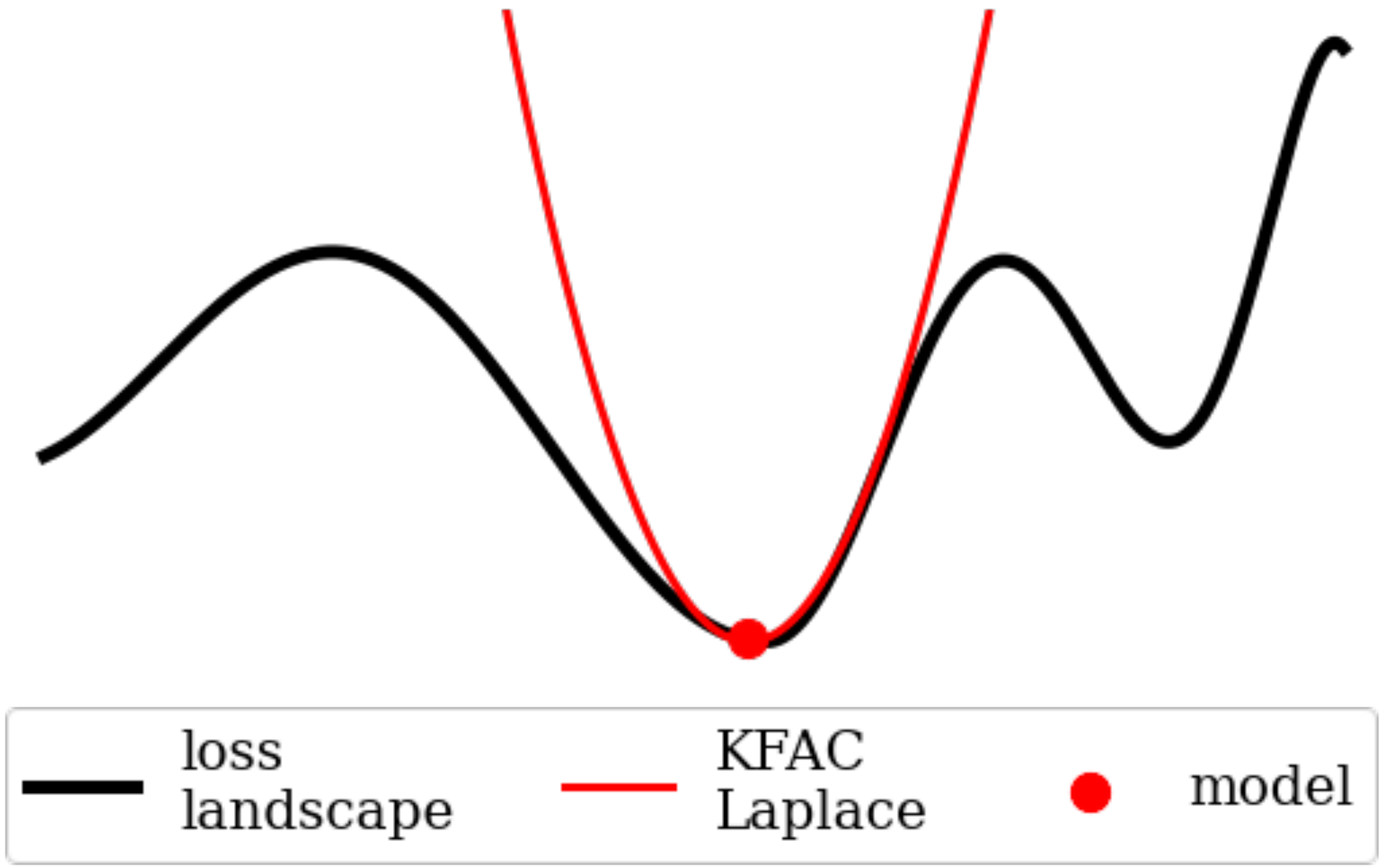
StarSSE results: ensembles



ResNet-50, CIFAR-100, BYOL self-supervised pre-training.

- **Optimal StarSSE** outperforms both **optimal SSE** and **Local DE**
- **Semi-local StarSSE** quality degrades less than **semi-local SSE**

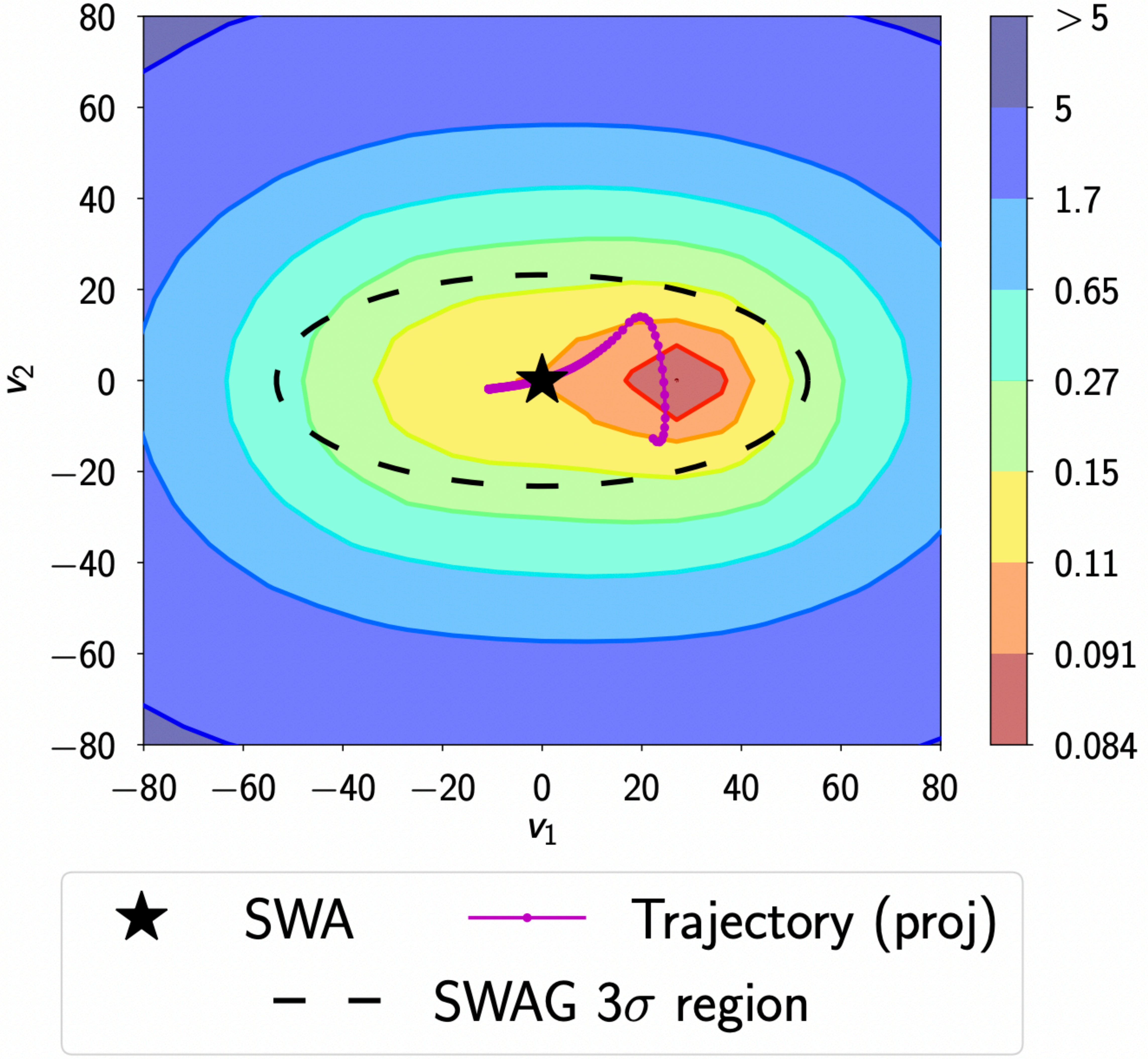
Non-cyclical local methods



KFAC Laplace (Ritter et al, 2018)

- Fit a Gaussian around a single trained model
- Kronecker factored approximation of Hessian matrix as covariance
- Sample new ensemble models from the Gaussian

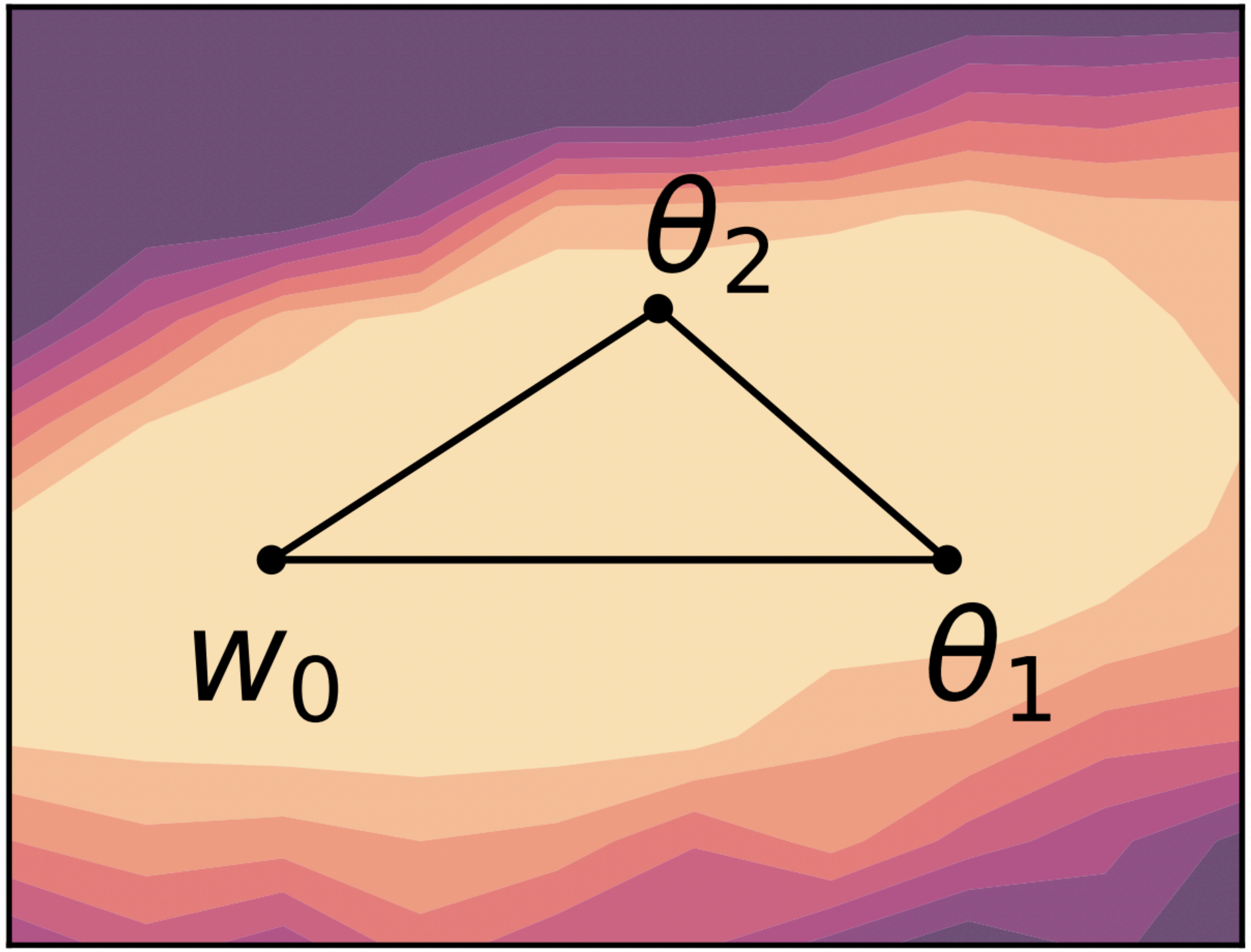
Non-cyclical local methods



SWAG (Maddox et al, 2019)

- Fit a Gaussian over models from training trajectory (SWA models)
- Requires additional epochs of training
- Sample new ensemble models from the Gaussian

Non-cyclical local methods



SPRO (Benton et al, 2021)

- Fit a simplex (e.g., a triangle) in the vicinity of a trained model
- Requires additional epochs of training
- Sample new ensemble models from the simplex

Non-cyclical local methods

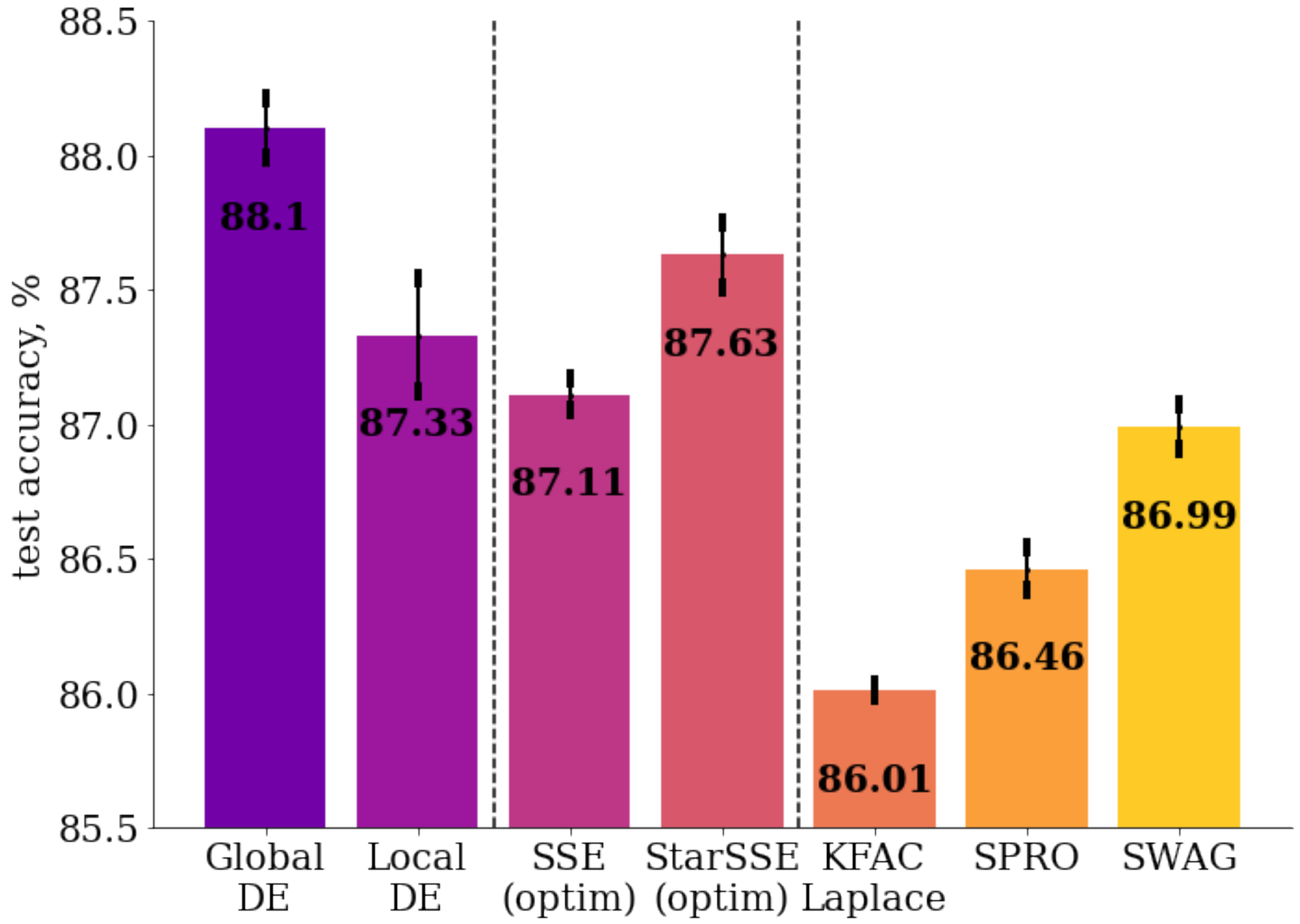
Comparison metrics:

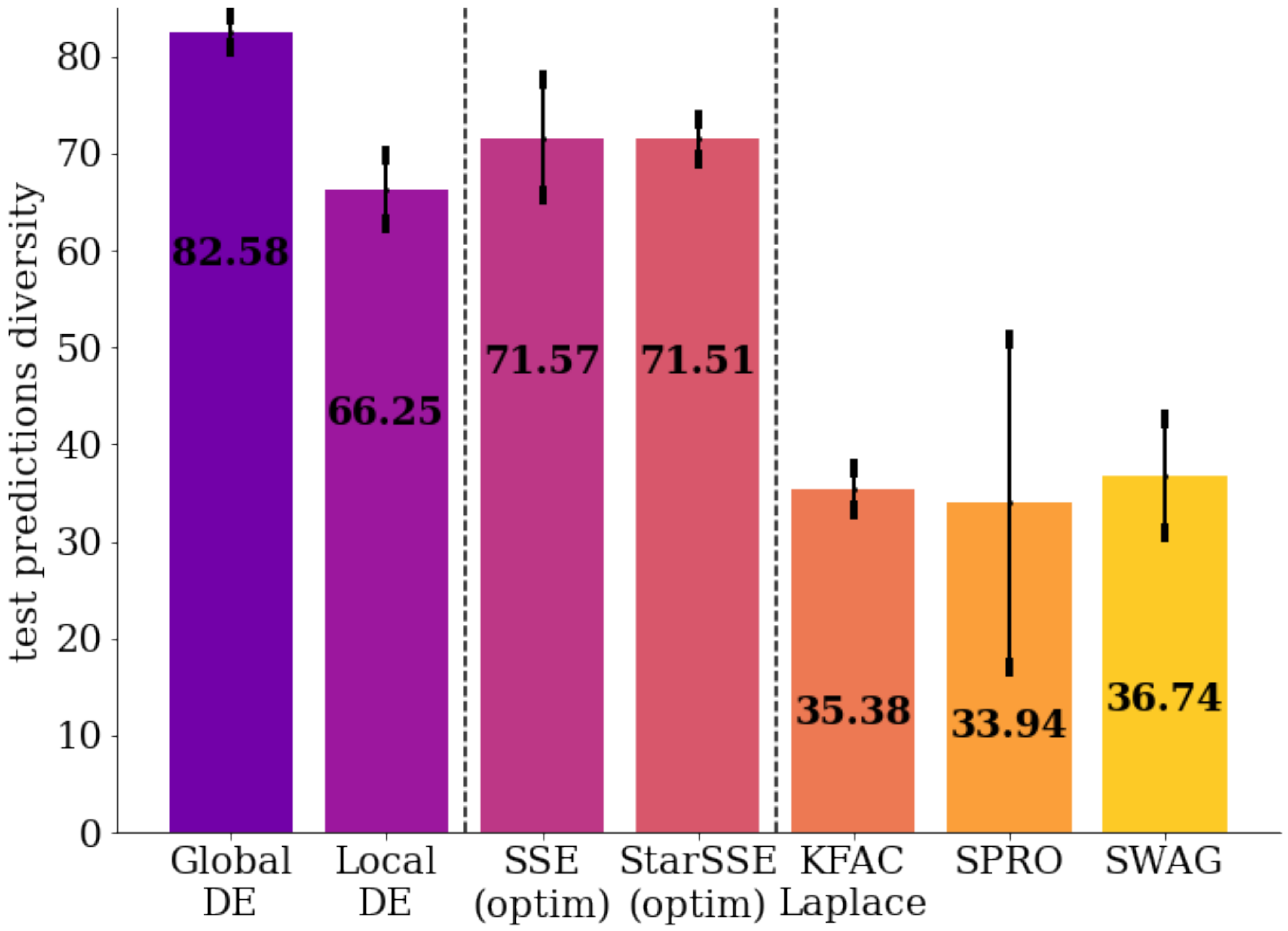
- test accuracy
- test prediction diversity:

$$diversity = 100 \cdot \mathbb{E}_{m_1 \neq m_2} \frac{\mathbb{E}_{images}\left[pred_1 \neq pred_2\right]}{\max(err_1, err_2)}$$

- m_i — model from the ensemble
- $pred_i$ — prediction of model m_i for a given image
- err_i — test error of model m_i

Non-cyclical local methods

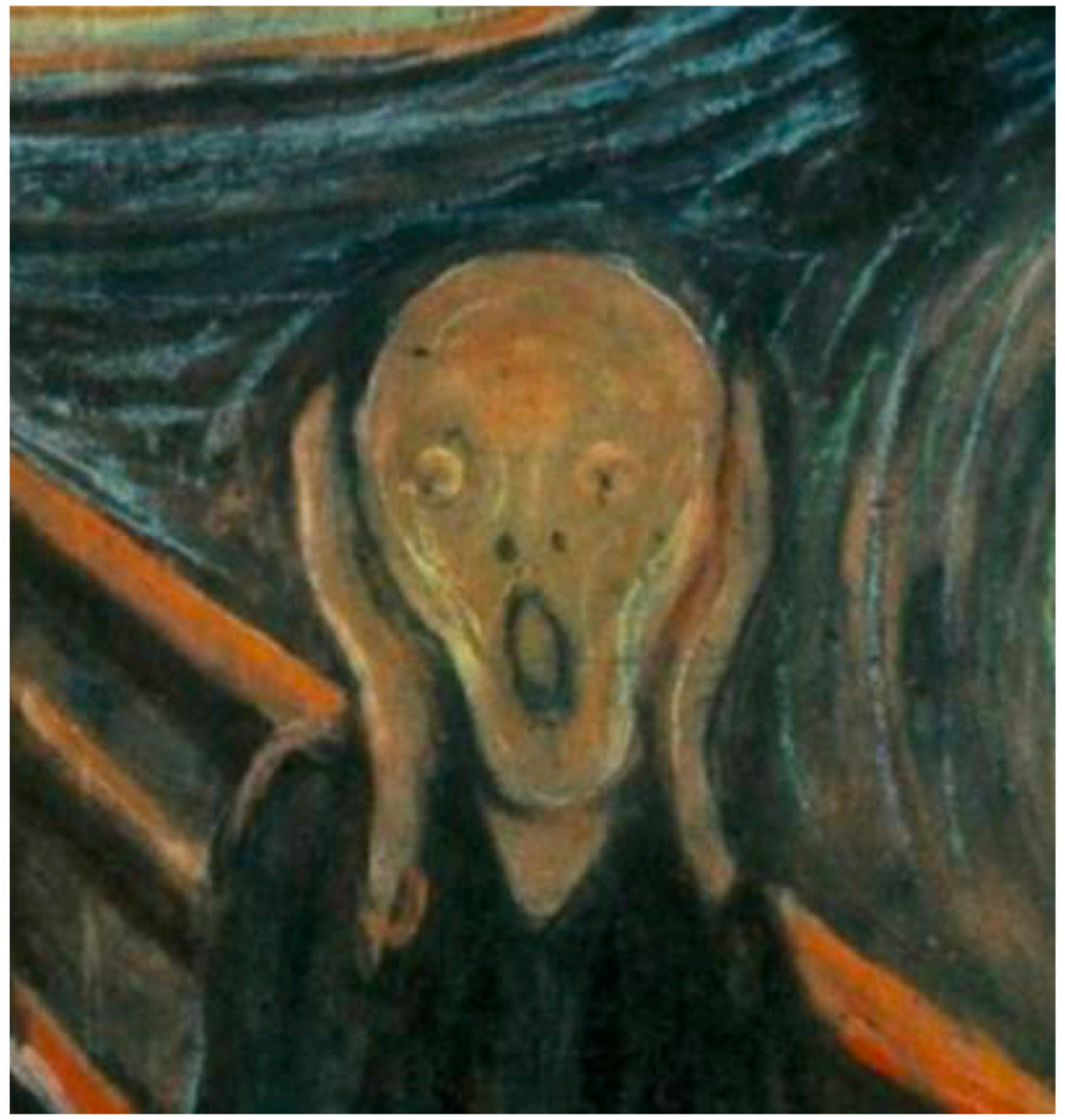




ResNet-50, CIFAR-100, BYOL self-supervised pre-training.

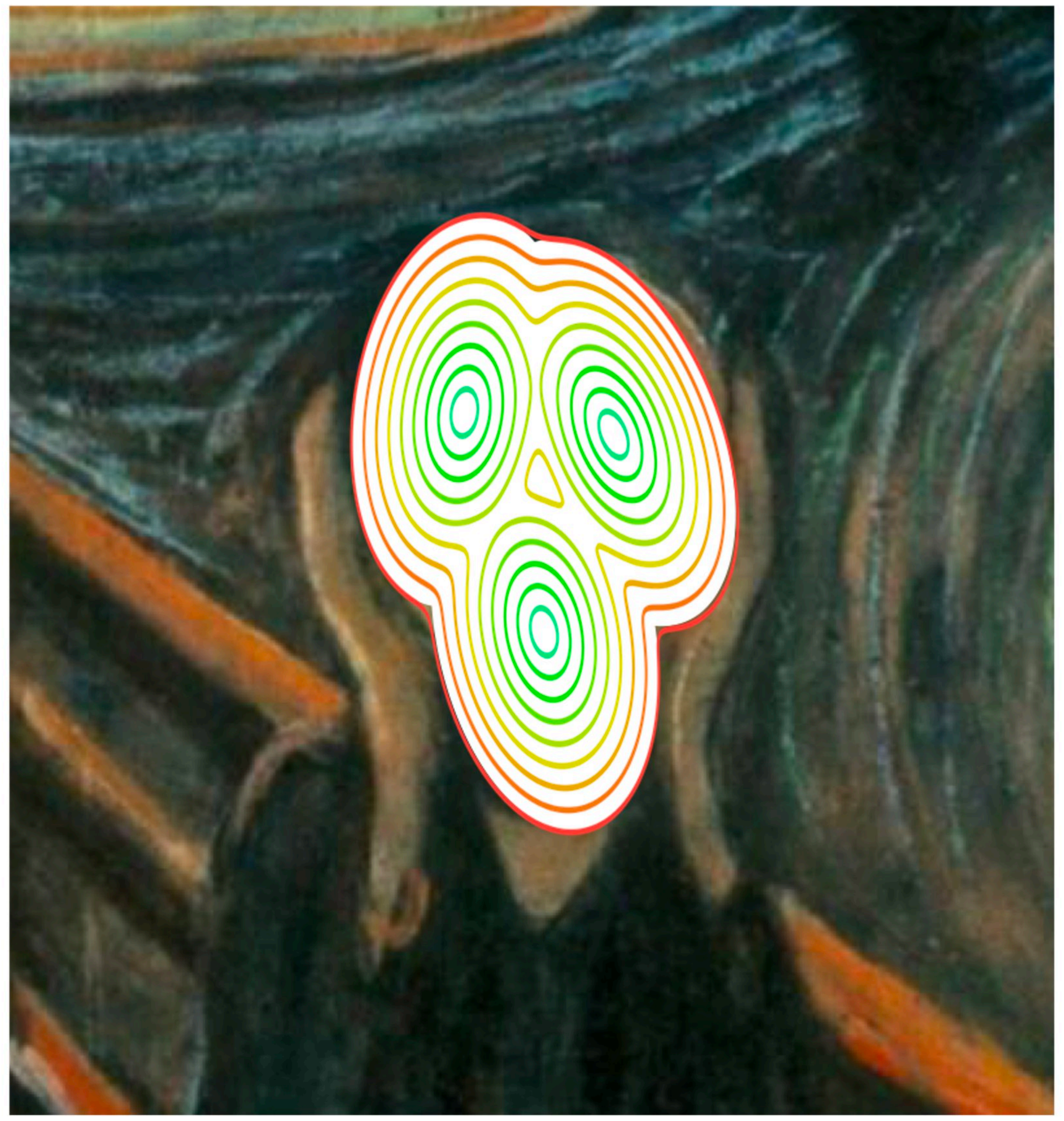
Feeling tired? Take a meme:

Usual researchers

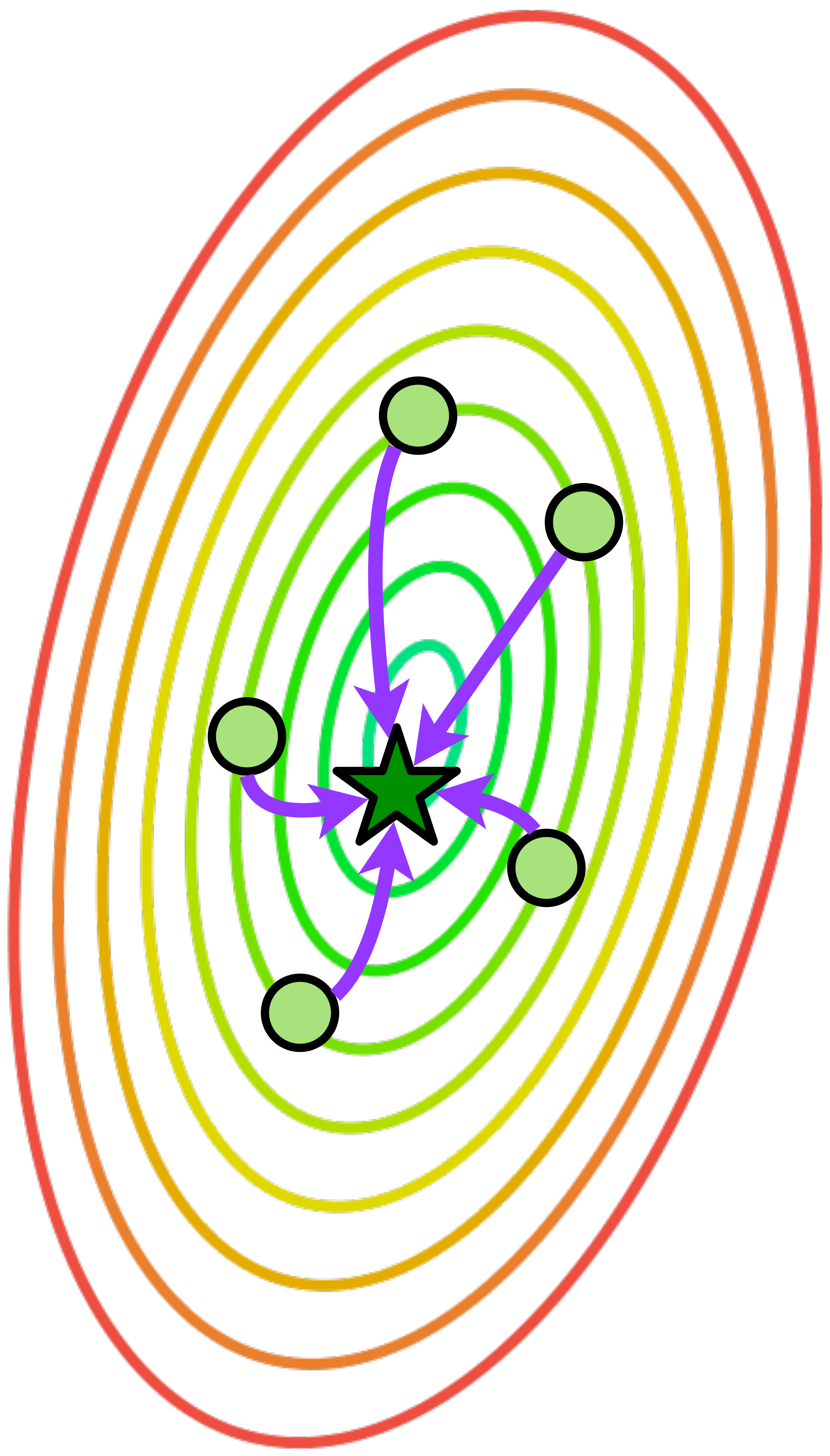


~~Listeners of this talk~~

Loss landscape researchers



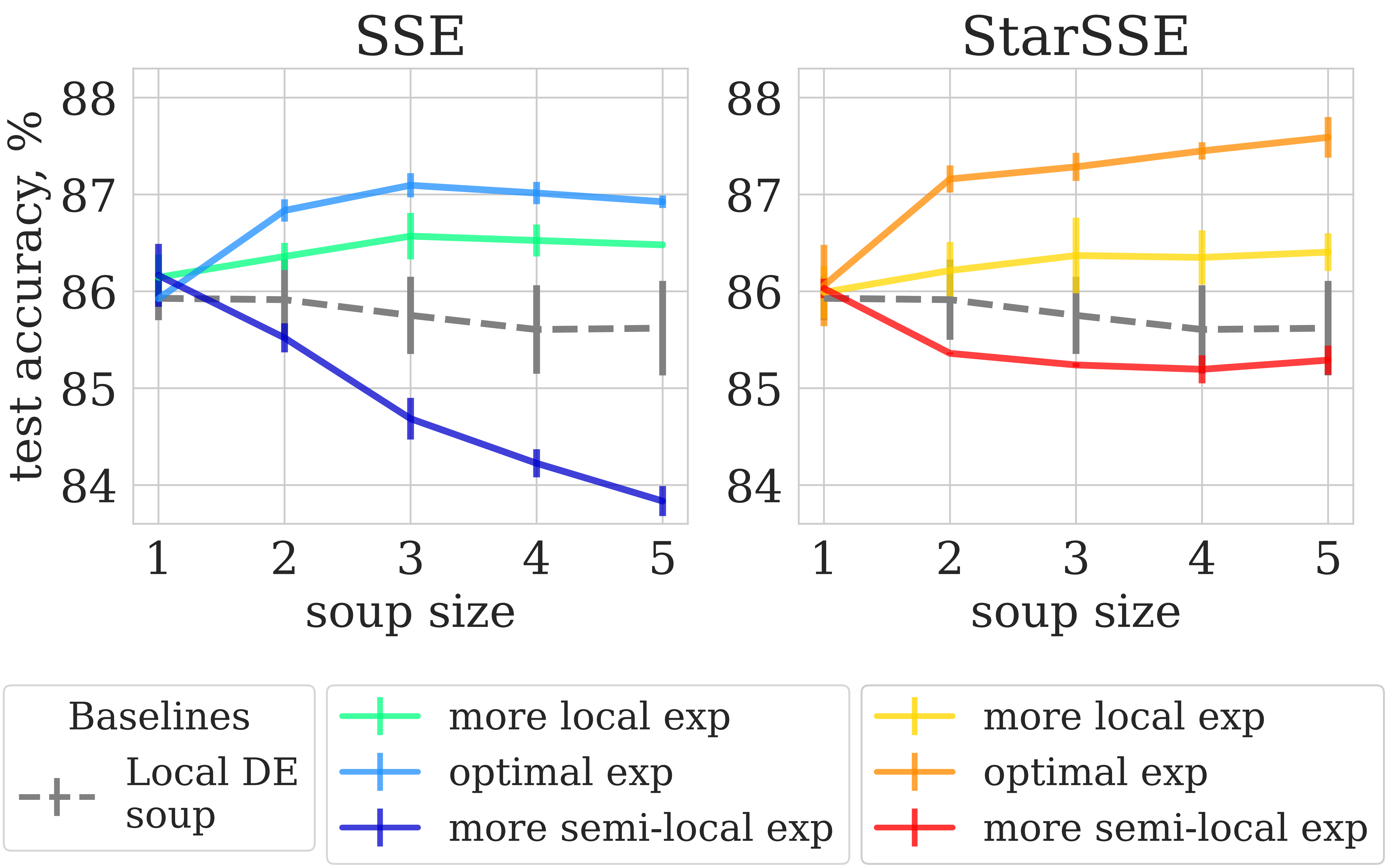
Model soups



By Wortsman et al, 2022

- Utilizing locality explicitly
- Average weights instead of predictions
- Faster inference
  (1 forward pass instead of N)
- Good OOD performance

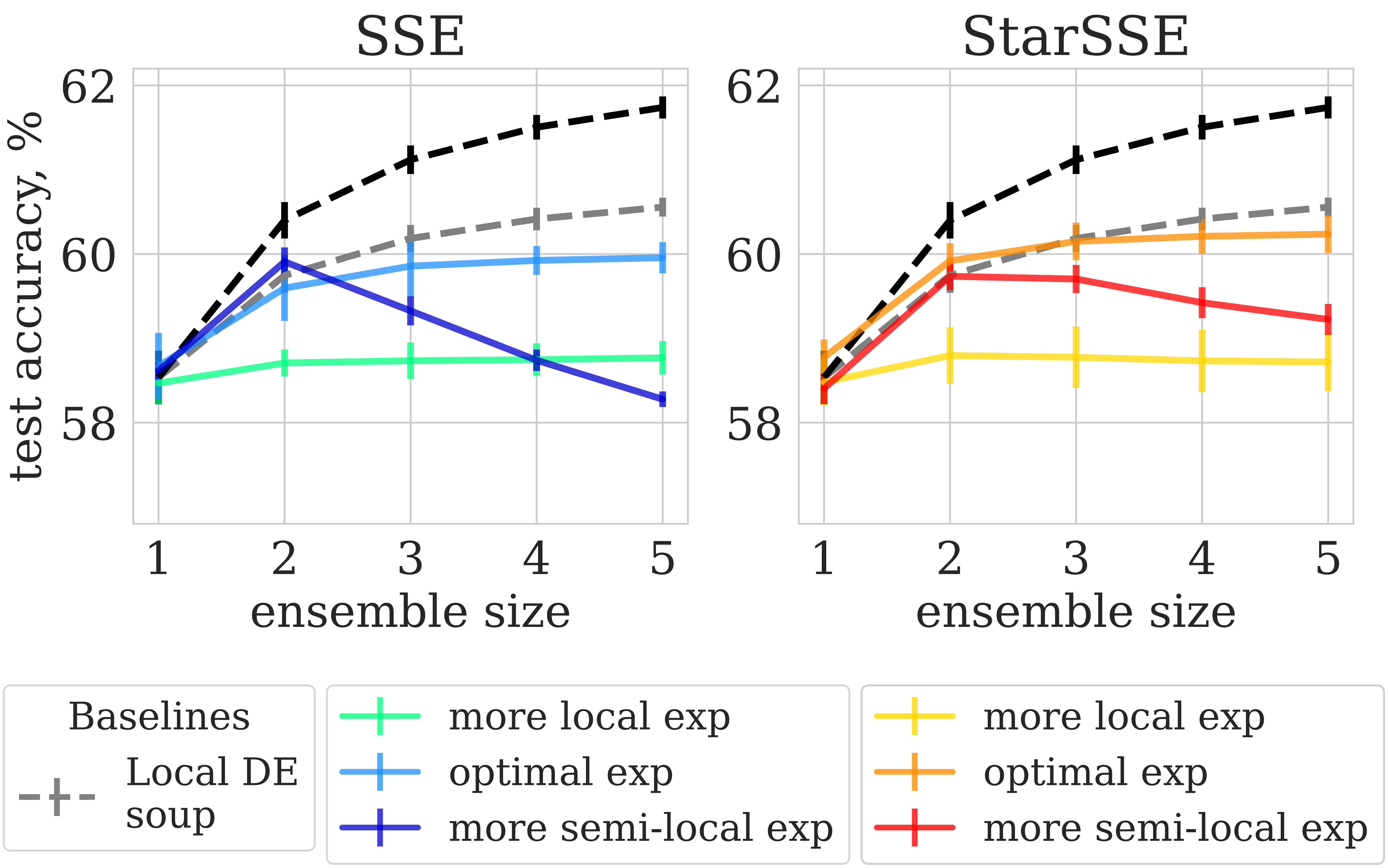
StarSSE results: model soups



ResNet-50, CIFAR-100, BYOL self-supervised pre-training.

- **Optimal StarSSE soup** outperforms both **optimal SSE soup** and **Local DE soup**
- StarSSE find models:
  - ✓ **more diverse** than Local DE and forms strong ensembles
  - ✓ located in a **more "convex" region** than Local DE and forms good soups

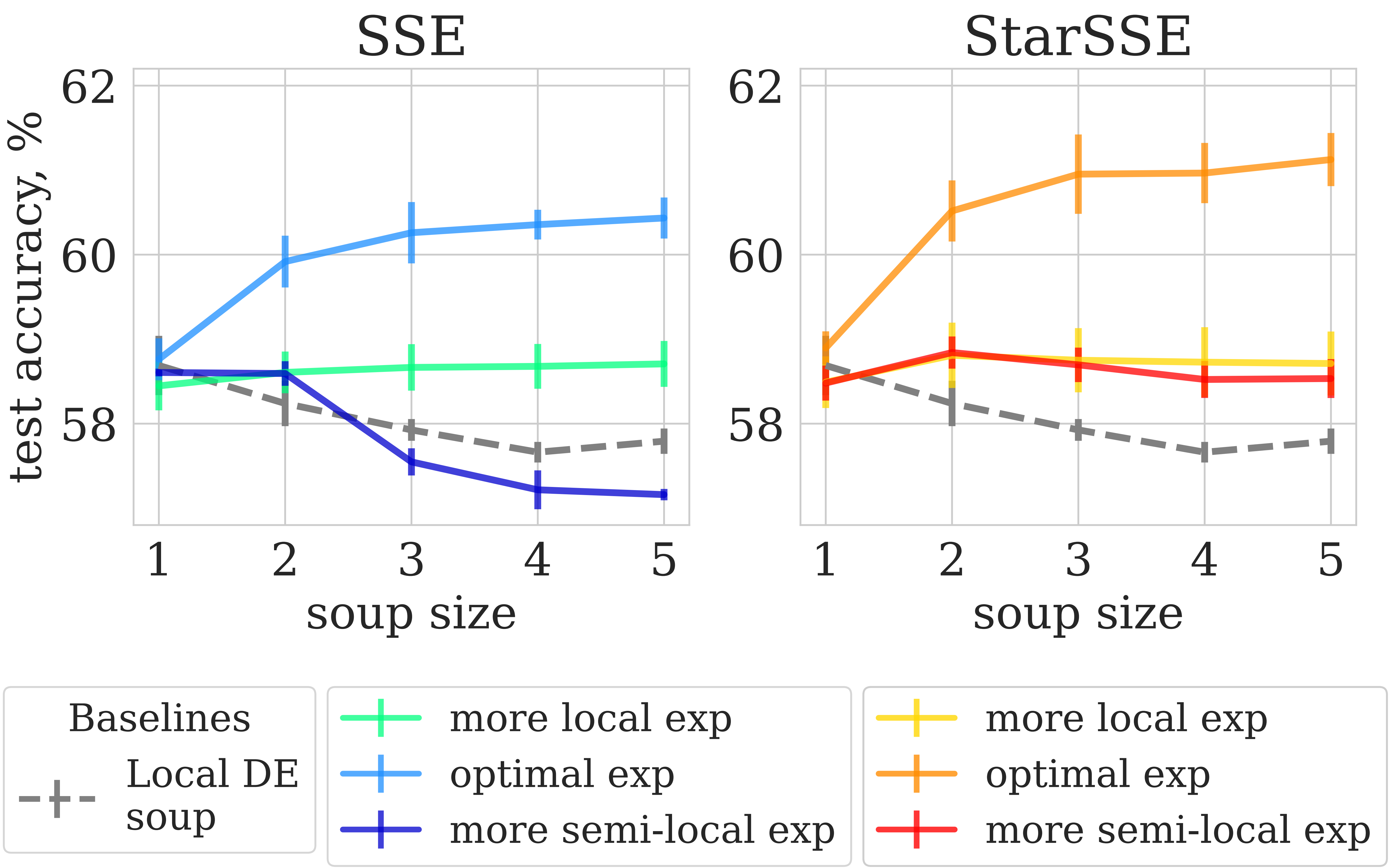
StarSSE results: OOD ensembles



ResNet-50, CIFAR-100C, BYOL self-supervised pre-training.

- CIFAR-100C: 19 synthetic corruptions, 5 severity values
- **Optimal StarSSE** and **optimal SSE** become inferior to **Local DE**
- Degradation of individual models quality is more pronounced on OOD data

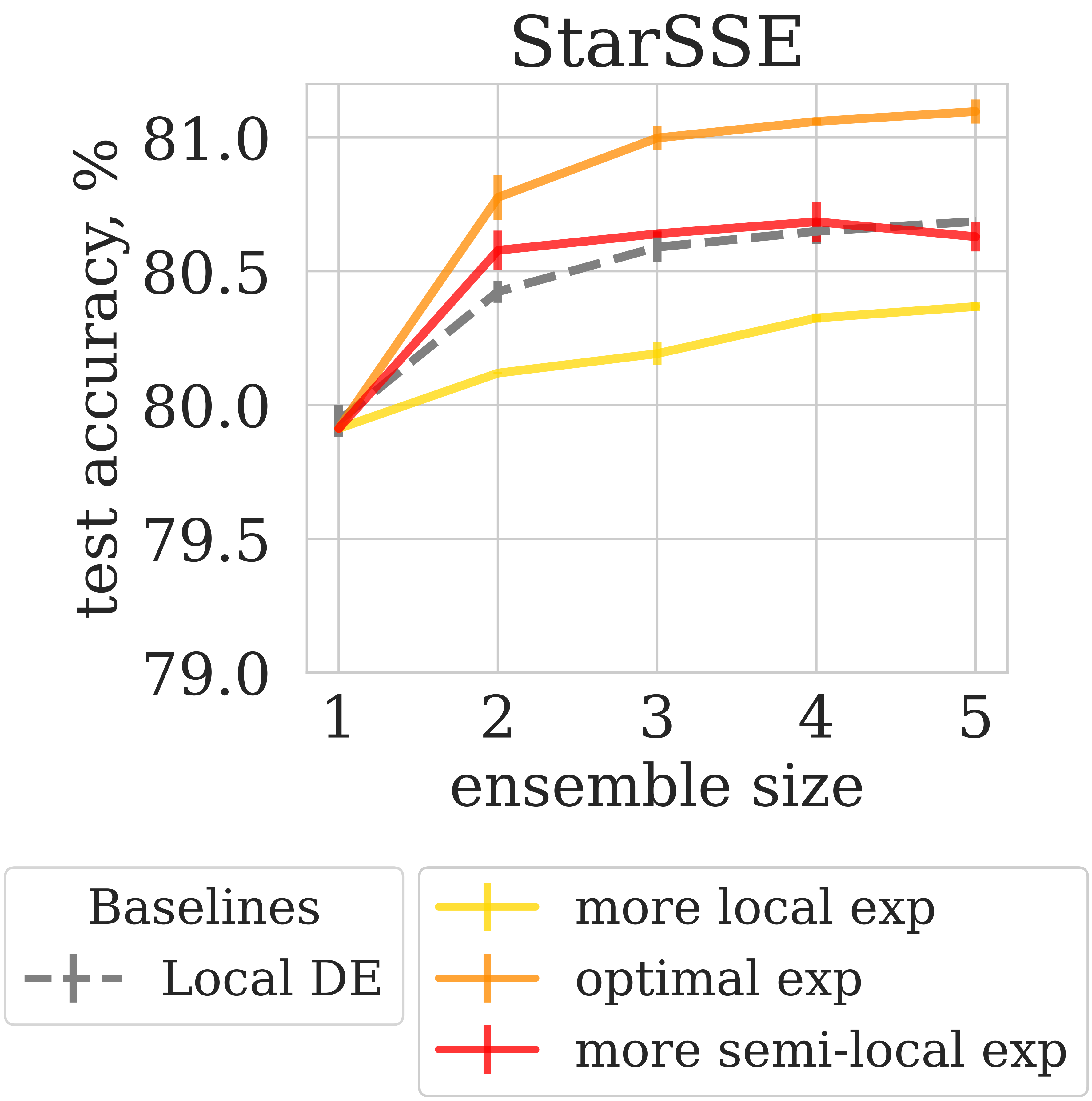
StarSSE results: OOD soups



ResNet-50, CIFAR-100C, BYOL self-supervised pre-training.

- CIFAR-100C: 19 synthetic corruptions, 5 severity values
- **Optimal StarSSE soup** has the best OOD performance

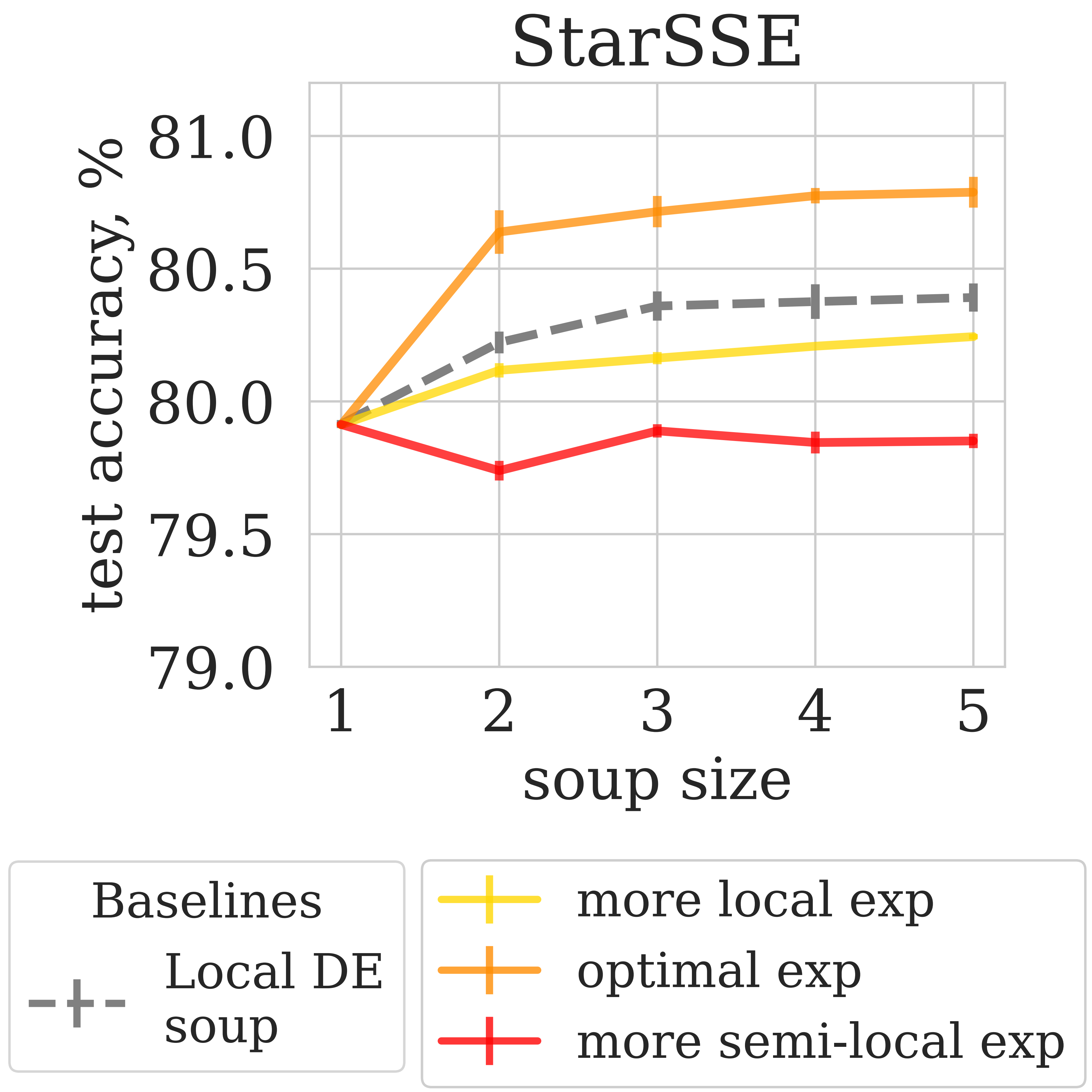
Large scale experiment: ensemble



StarSSE works in a more practical setup as well:

- ViT-B/32 architecture
- CLIP pre-training
- ImageNet fine-tuning

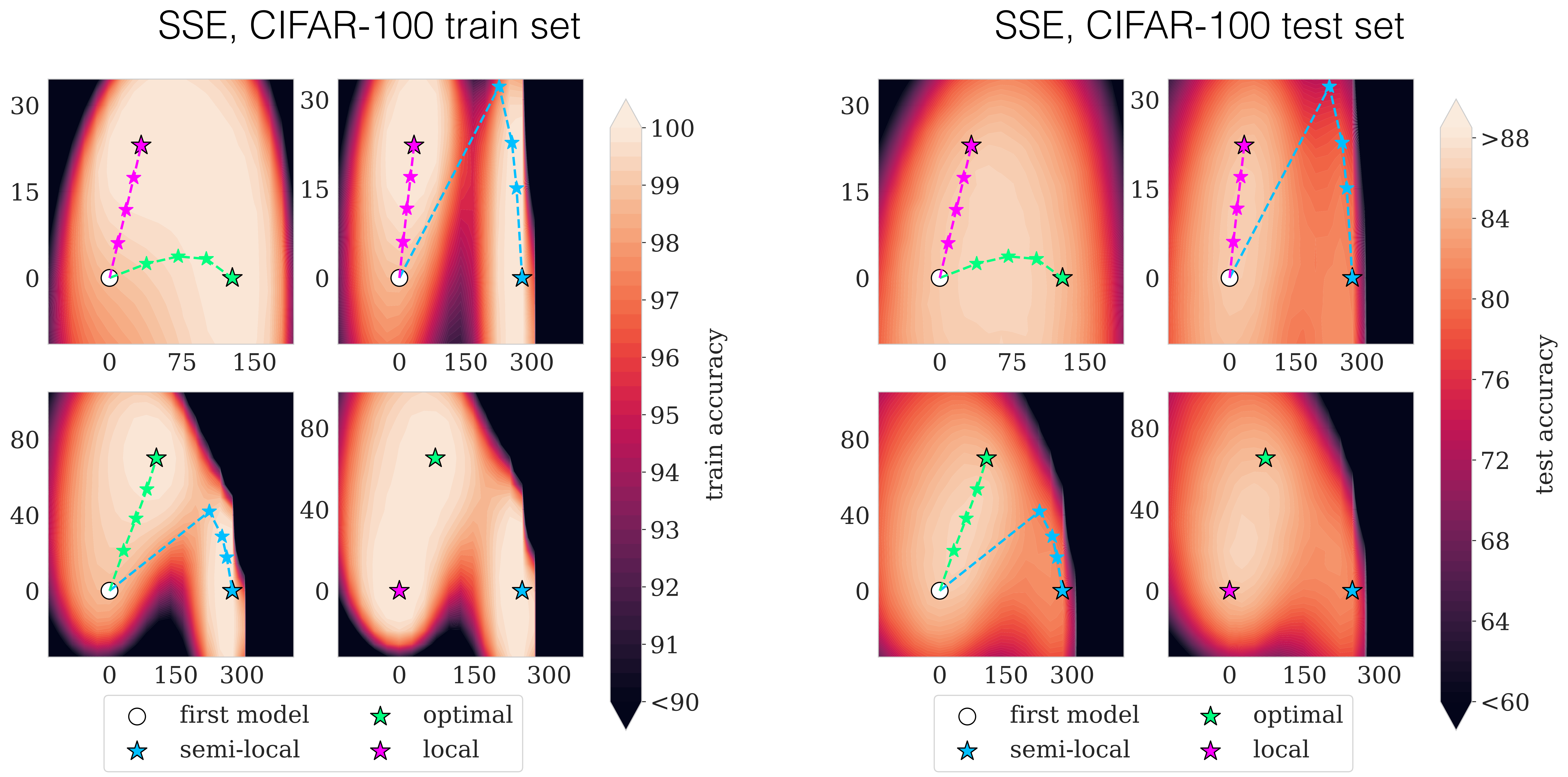
Large scale experiment: model soup



StarSSE works in a more practical setup as well:

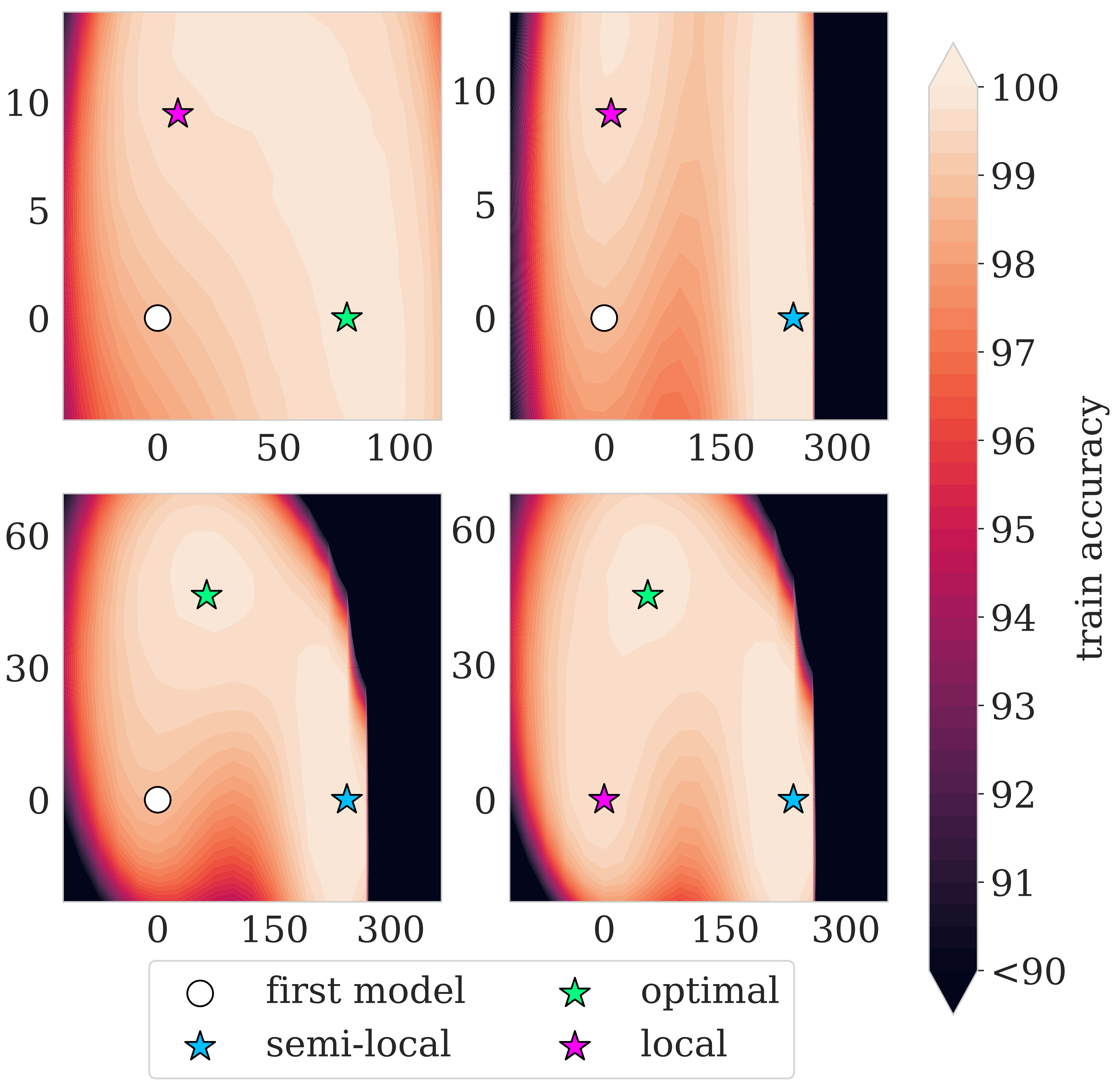
- ViT-B/32 architecture
- CLIP pre-training
- ImageNet fine-tuning

2D loss landscape visualization

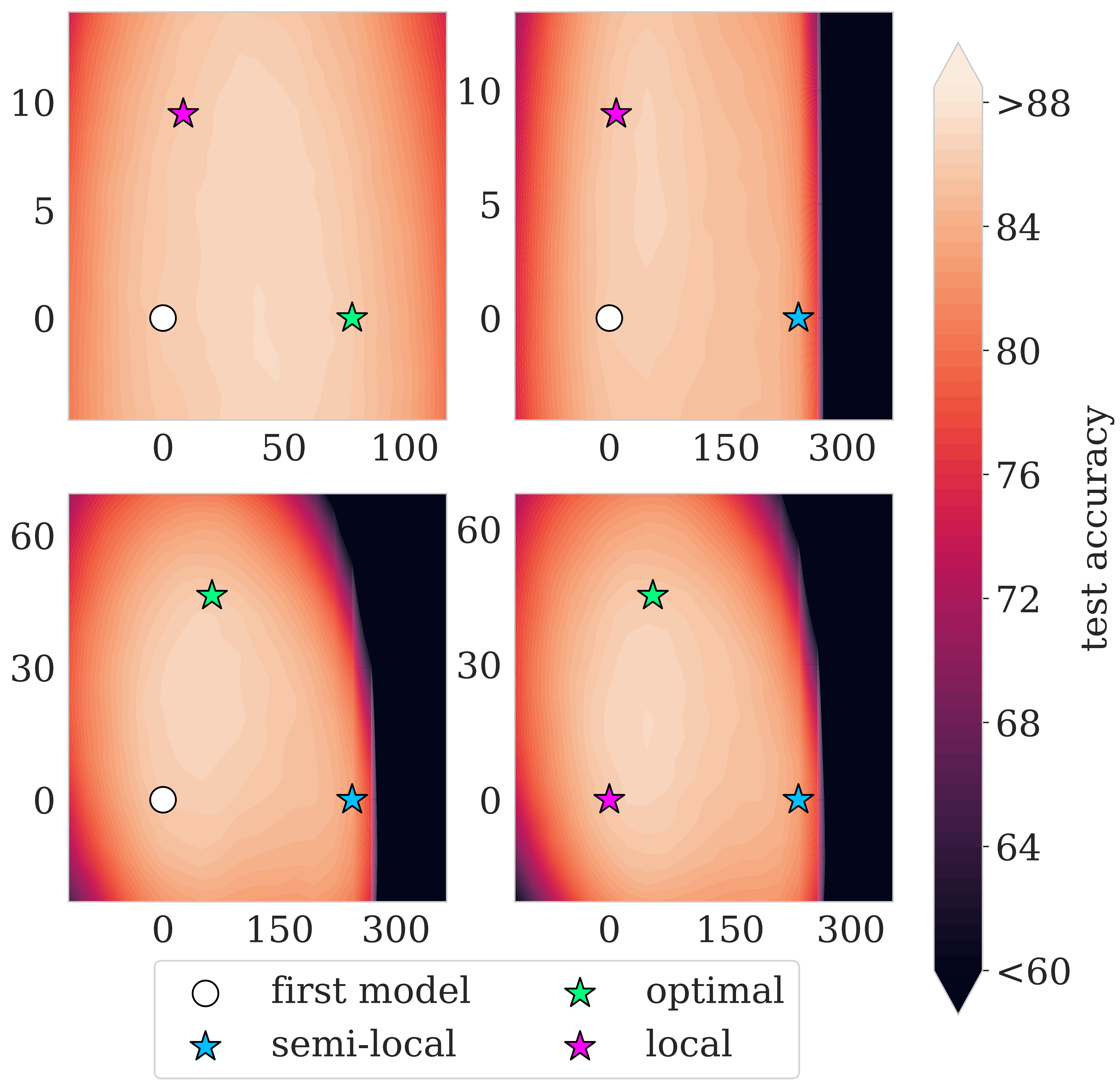


2D loss landscape visualization

StarSSE, CIFAR-100 train set



StarSSE, CIFAR-100 test set



Conclusion

- SSE does not close the gap between Local and Global DE
  ✓ local behavior — high accuracy ensembles
  ✗ semi-local behavior — degradation of models quality

- StarSSE — parallel modification of SSE
  ✓ better suits specific of transfer learning
  ✓ outperforms both SSE and Local DE
  ✓ strong model soups (especially on OOD!)

- Additional results: other datasets, model diversification analysis

Paper: https://arxiv.org/abs/2303.03374
Code: https://github.com/isadrtdinov/ens-for-transfer

p(B|A)yesgroup