COLUMBIA | Zuckerman Institute

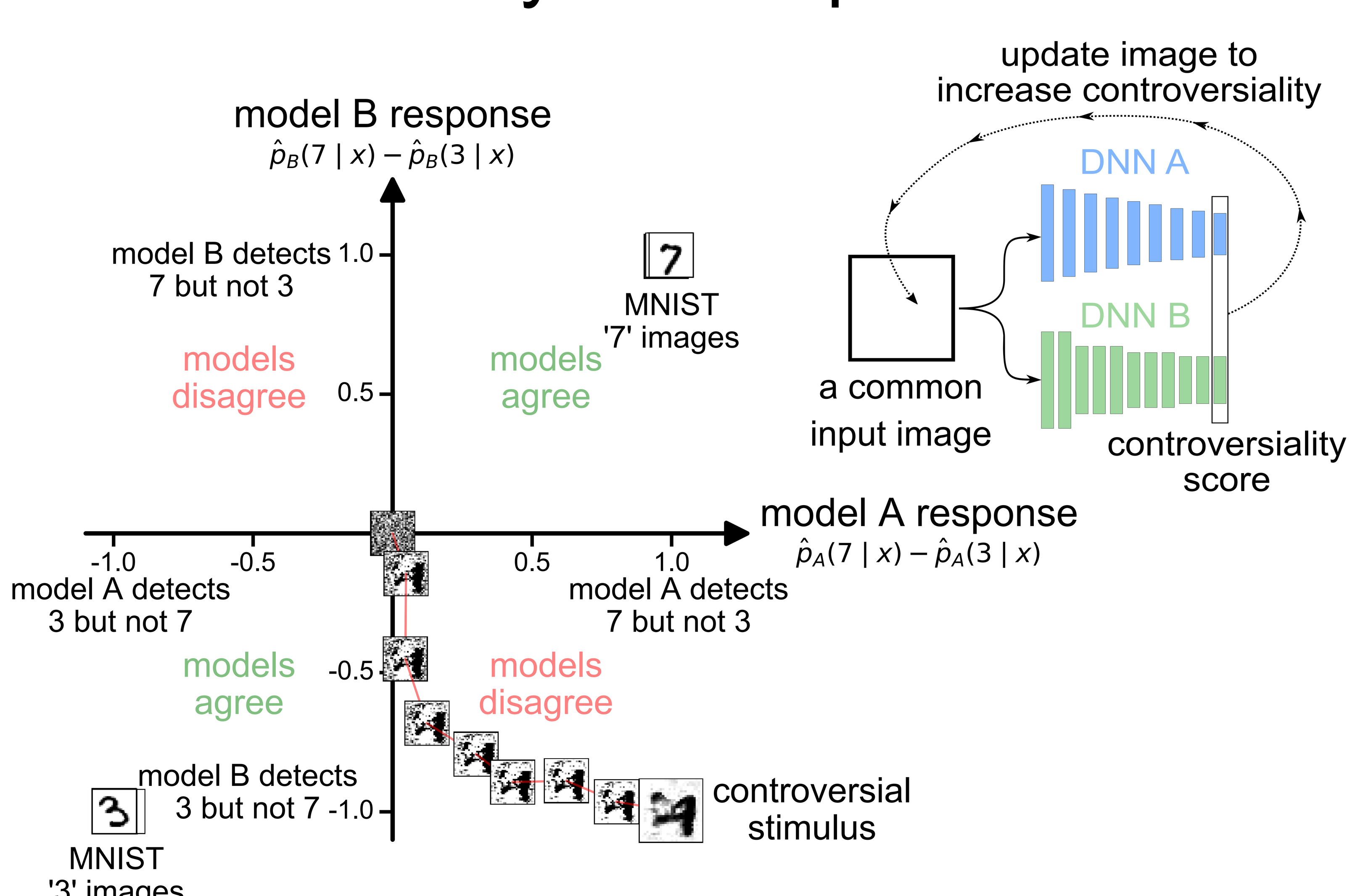
Adjudicating between deep neural network models of biological vision with controversial stimuli

Tal Golan, Prashant C. Raju, Nikolaus Kriegeskorte

To efficiently adjudicate between deep neural network models of biological vision, we must devise testing conditions in which different models make different predictions.

We suggest using **controversial stimuli**: images synthesized to make different models disagree.

Stimulus synthesis procedure



model A detects digit d_a — but not d_b

$$c_{A,B}^{d_a,d_b}(x) = \min\{\hat{p}_A(d_a \mid x), 1-\hat{p}_A(d_b \mid x), \hat{p}_B(d_b \mid x), 1-\hat{p}_B(d_a \mid x)\}$$

the controversiality of image x with respect to models A and B and digits d_a and d_b

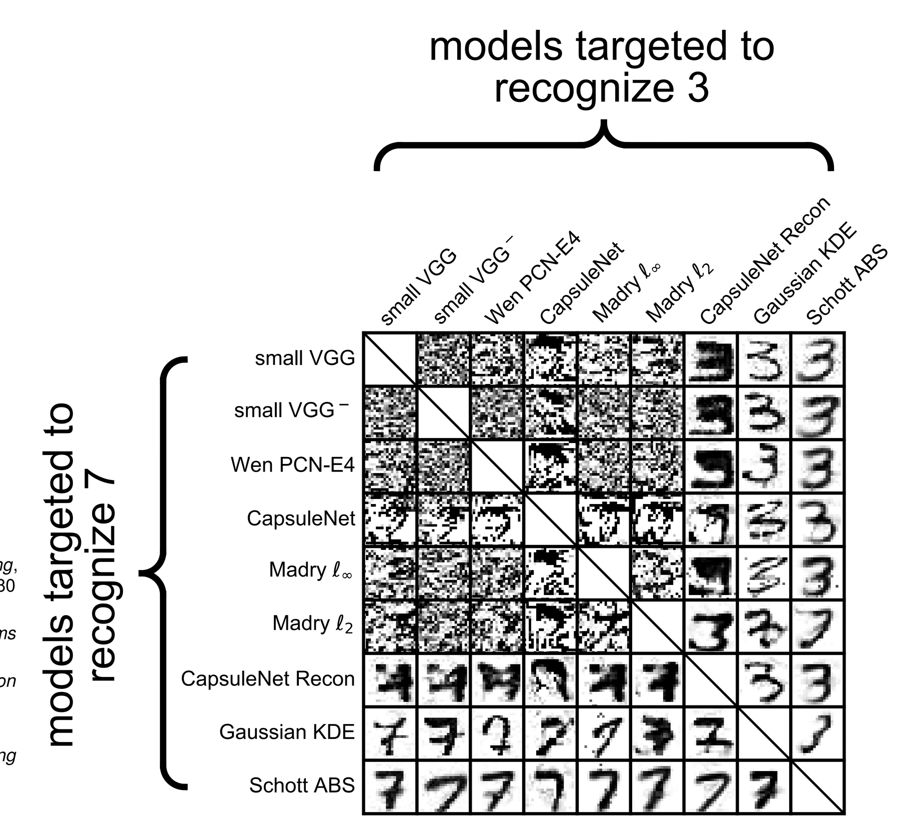
model B detects digit d_b — but not d_a

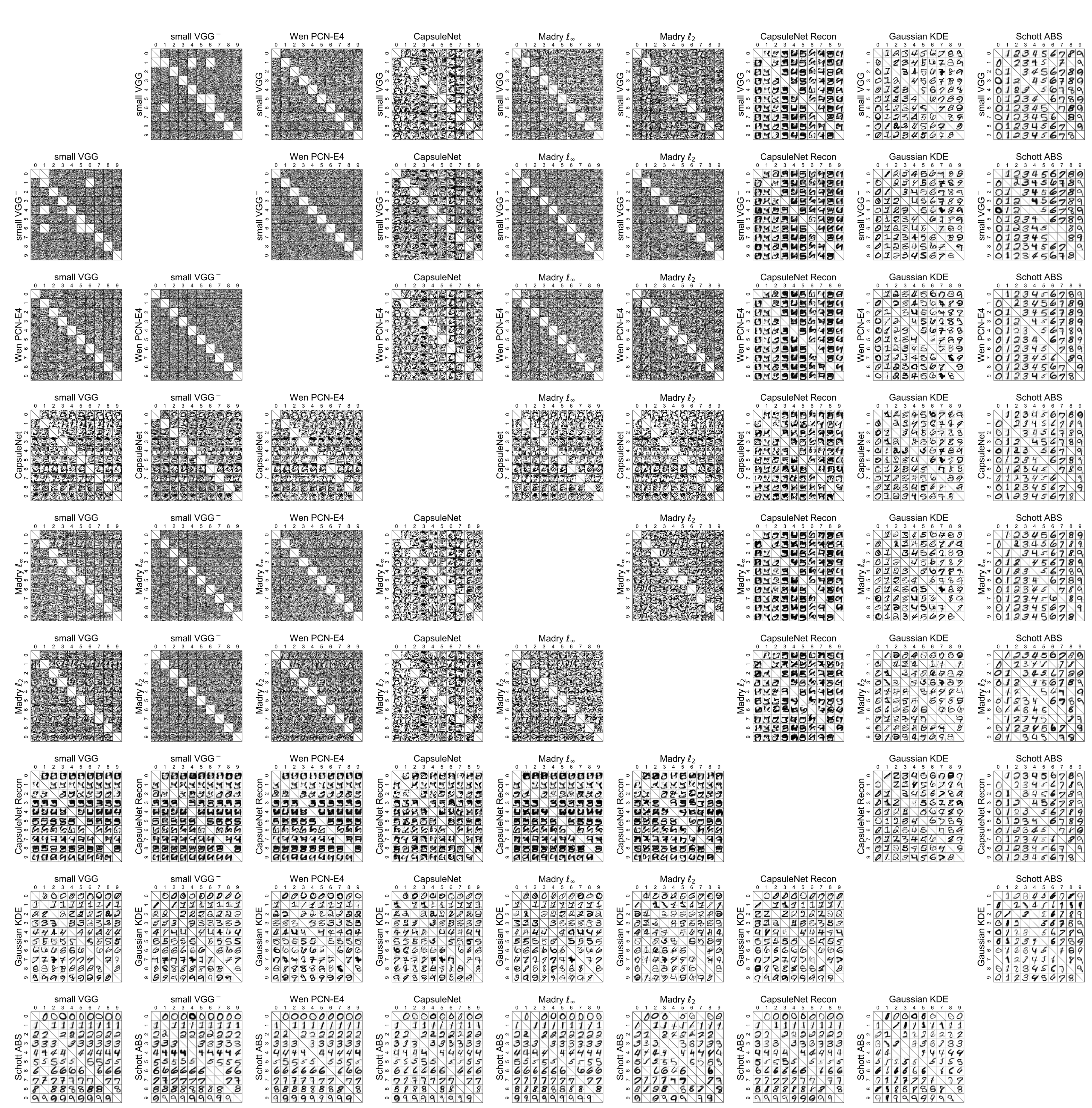
Controversial stimuli for MNIST-classifying DNNs

Tested models

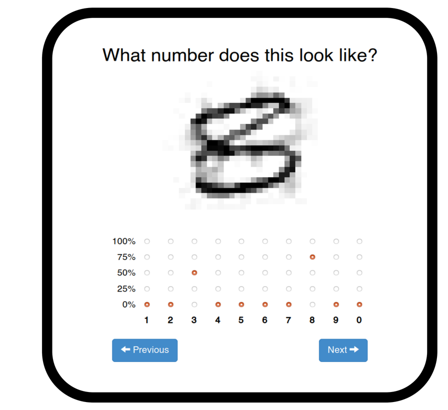
model family	model	error
discriminative feedforward	small VGG [1]*	0.47%
	small VGG⁻ [1]*	0.59%
discriminative recurrent	Wen PCN-E4 [2]	0.42%
	CapsuleNet [3]	0.24%
adversarially trained	Madry ℓ_∞ [4] ($\epsilon = 0.3$)	1.47%
	Madry ℓ_2 [4] ($\epsilon = 2$)	1.07%
reconstruction-based	CapsuleNet Recon [5]*	0.29%
generative	Gaussian KDE	3.21%
	Schott ABS [6]	1.00%

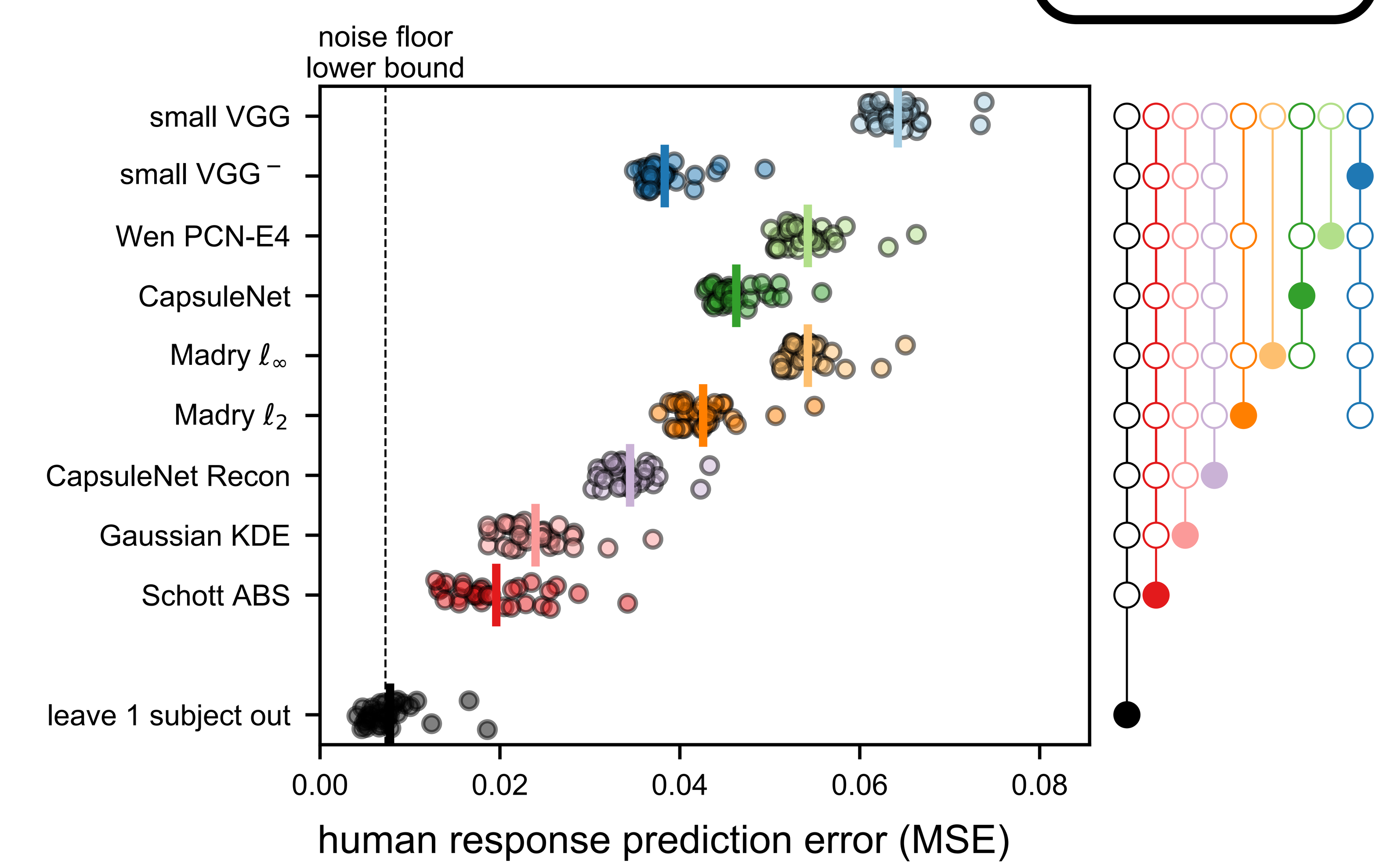
[1] K. Simonyan, A. Zisserman, *arXiv preprint arXiv:1409.1556* (2014).
[2] H. Wen, *et al.*, *Proceedings of the 35th International Conference on Machine Learning*, J. Dy, A. Krause, eds. (PMLR, Stockholmsmässan, Stockholm Sweden, 2018), vol. 80 , pp. 5266–5275.
[3] S. Sabour, N. Frosst, G. E. Hinton, *Advances in Neural Information Processing Systems 30*, I. Guyon, *et al.*, eds. (Curran Associates, Inc., 2017), pp. 3856–3866.
[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, *International Conference on Learning Representations* (2018).
[5] Y. Qin, *et al.*, *CoRR* abs/1907.02957 (2019).
[6] L. Schott, J. Rauber, M. Bethge, W. Brendel, *International Conference on Learning Representations* (2019).





Human experiment: predicting human ratings by DNN outputs





➤ Controversial stimuli allow to efficiently compare DNN models of vision.

➤ Each controversial stimulus must be an adversarial example for at least one model. This does not hinge on presumed invisibility of ℓ_p-bounded perturbations.

➤ For MNIST, class-conditional generative models predict human perception better than discriminative models.

See our full arXiv preprint!
https://arxiv.org/abs/1911.09288