# A large-scale examination of inductive biases shaping high-level visual representation in brains and machines


**Paper Presentation** -

**Harshvardhan Singh | 2022112004**

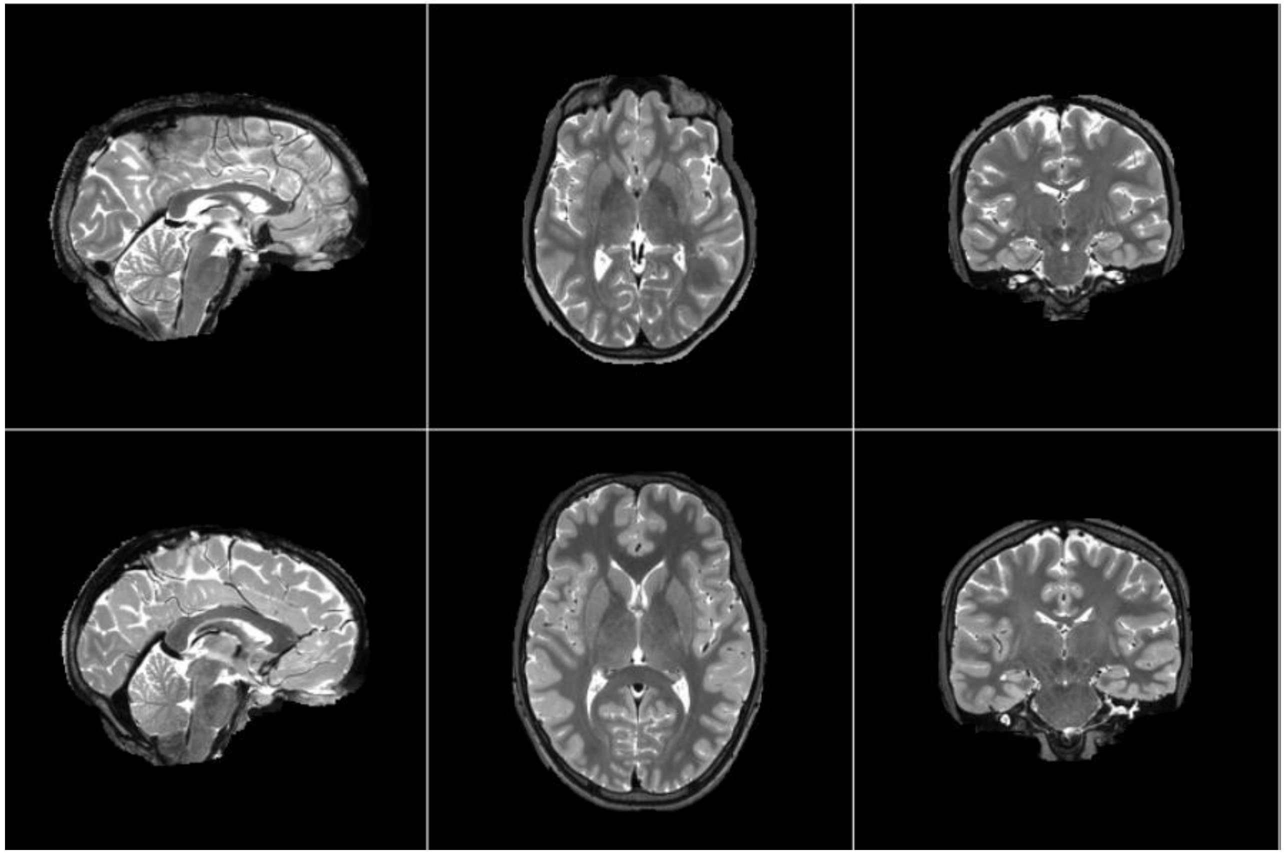
**Soham Vaishnav | 2022112002**

Group ID - **8**

Authors: **Conwell, C., Prince, J.S., Kay, K.N. et al.**

# Table of Contents

- **Problem Statement**
- **Related Works**
  - Motivation
- **Methods**
- **Results**
- **Discussion**
- **Novelty**
- **Conclusion**
  - Limitations
  - Future directions

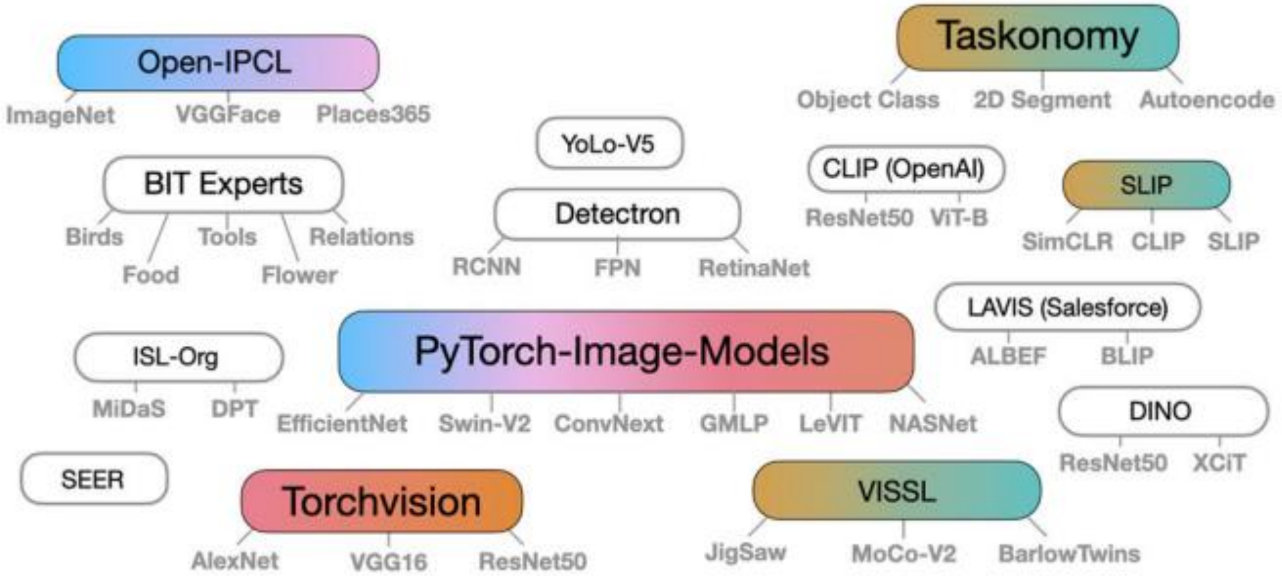


# Problem Statement

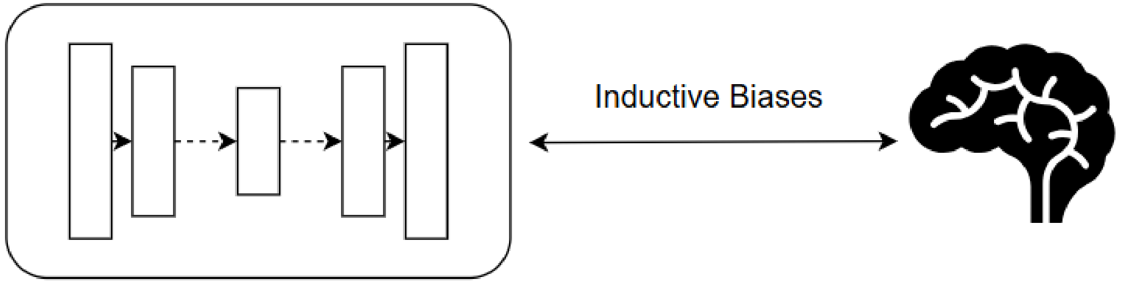
- Lacking computational clarity on the later stages of visual ventral stream
- Boom in number of vision models
  - Provides traction for conducting empirical testing for decoding high-level visual representations

- Strong correlations between internal latent space and hierarchical representations of DNNs, and structure of responses in biological systems
- DNN models designed for canonical CV tasks

Demands revamping in the way we approach alignment problems!



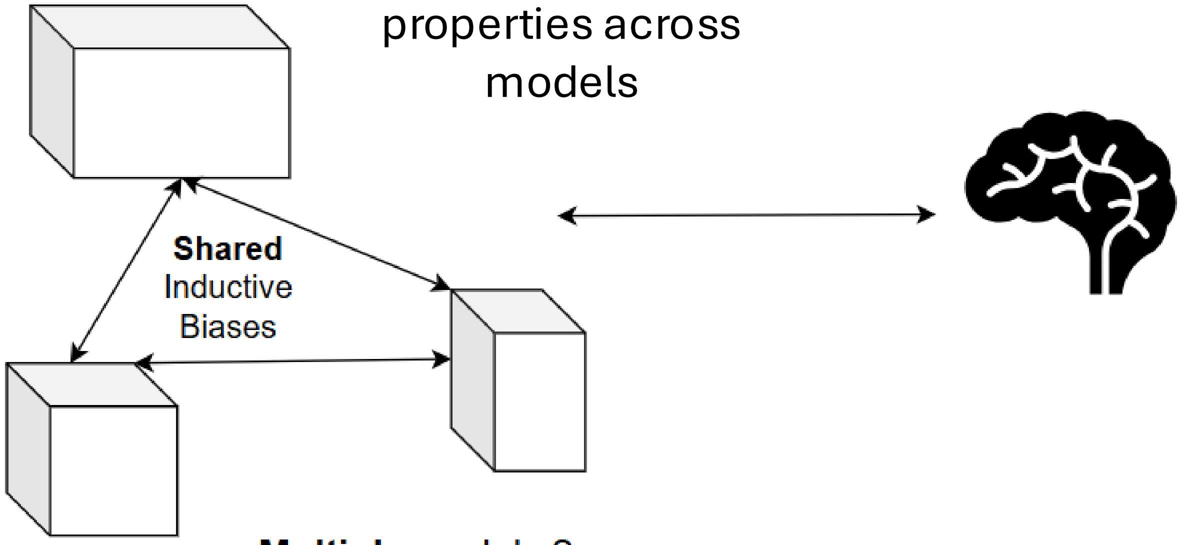
# Makes us ask two questions:




A **single** model ?

**Question 1:** single
go-to model for brain

**Question 2:** shared
properties across
models




**Multiple** models ?

# For **Question 1 …**

ALGONAUTS PROJECT

SENSORIUM 2023

Welcome to the Sensorium 2023 Competition!

Brain-Score

Current platforms to promote building and benchmarking of a single model that can work as closely as brain does

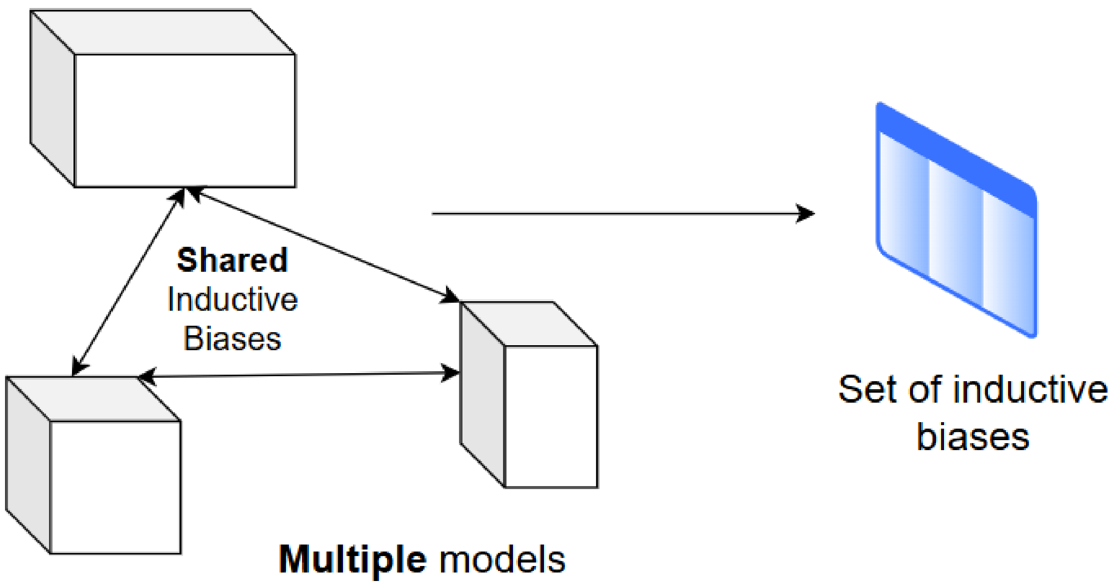
But for **Question 2 ….** is what this paper aims to explore

**Premise** of the paper:

Different DNNs learn high level representations based on their:

1. Architecture
2. Training Data
3. Objective

and other set of hyperparameters …




**Aim**: Comparing various models to determine the **set of inductive biases** that contribute to **most brain-like predictions**

Important to Note:

1. Models **not competing** to be the **best architectural** replication of the brain

2. Models simply **considered as visual representation learners**

3. Competition used to **derive** which **set of features** affects the learning process **the most**

# Related Works

# Diverse Deep Neural Networks All Predict Human Inferior Temporal Cortex Well, After Training and Fitting

Compared early computer vision models – set of 9 classical DNNs based on architecture

E.g. networks – **AlexNet**, **VGG**, **ResNet18**, etc.

Played around with features – did re-mixing and re-weighting

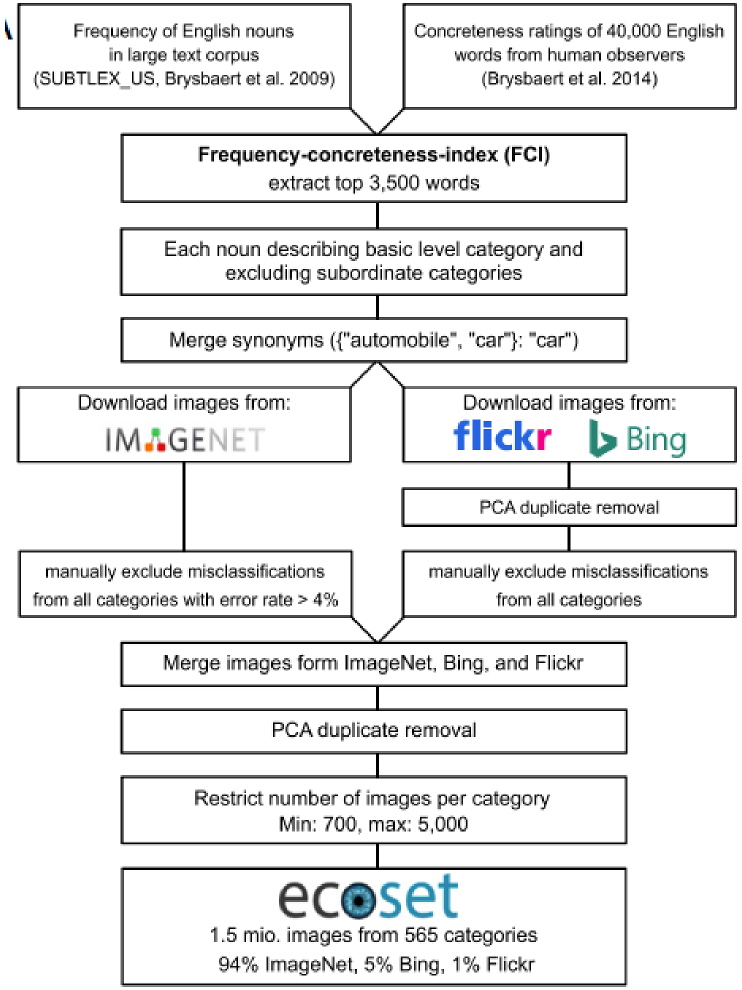
# An ecologically motivated image dataset for deep learning yields better models of human vision

Compared performance of multiple instances of two models:

**AlexNet** and **vNet** based on the dataset

Dataset created called **EcoSet –** had less erroneously labelled images and more ecologically relevant images

Better prediction of human high-level visual cortex representations



# Human Alignment of Neural Network Representations

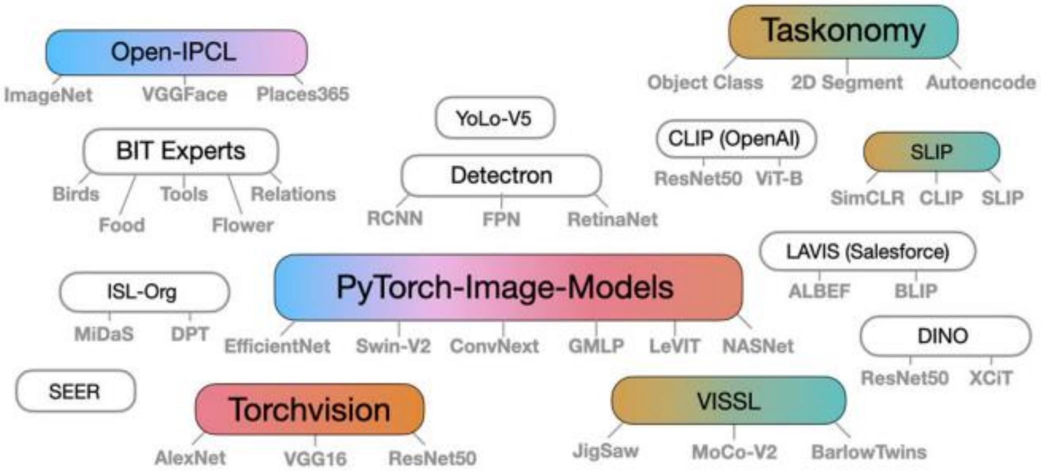
- Studied alignment of recently developed model architectures with human visual behaviour
- Models used include transformer-based architectures – or broadly put, attention-based mechanisms
- Also found that larger and more diverse datasets produce better alignment of judgement as produced in human visual behaviour

# Methods

# Model Selection

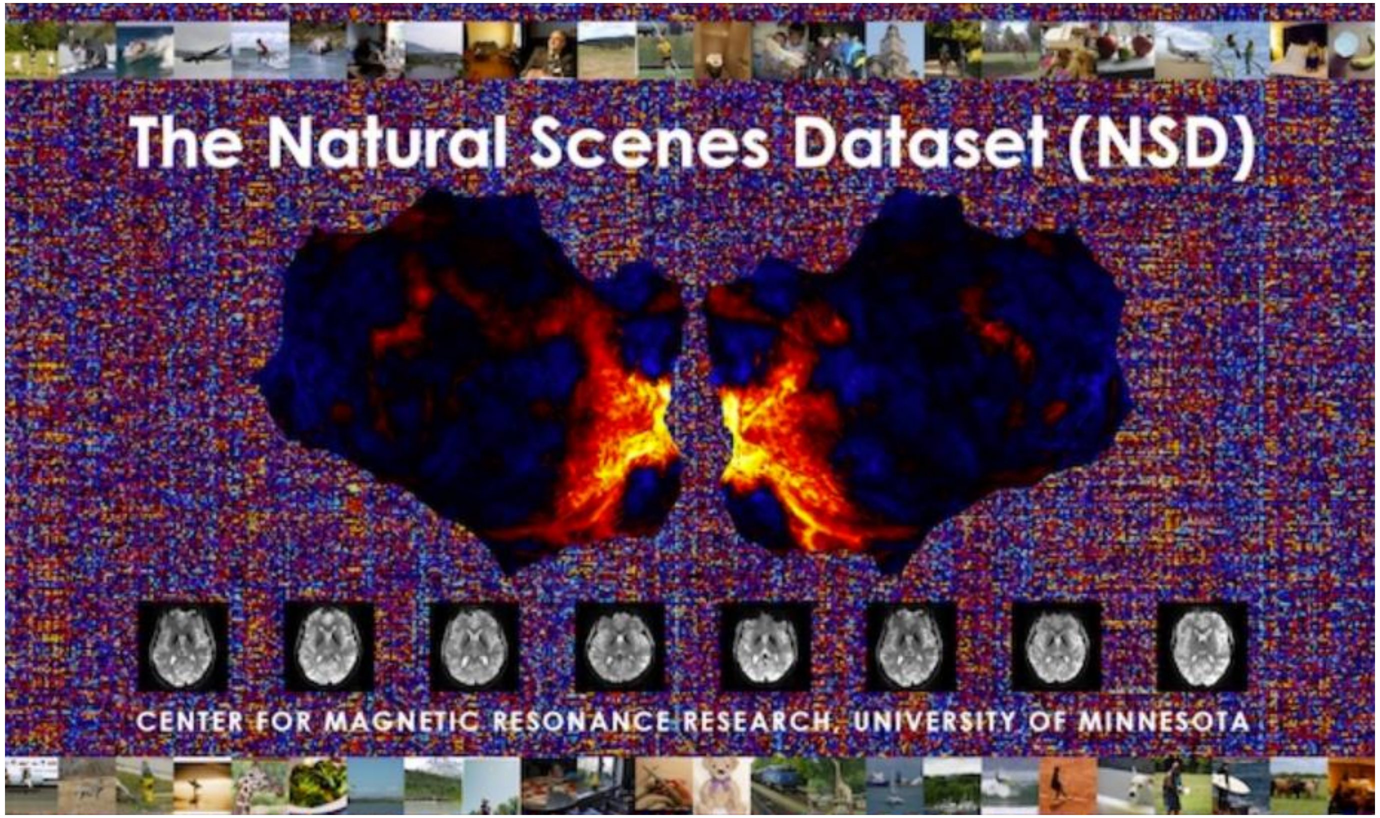
- 224 distinct models – 160 trained and 64 random init.
- Models taken from:
  - **Torchvision** (PyTorch) model zoo
  - **Pytorch-Image-Models** (timm) library
  - **VISSL** (self-supervised) model zoo
  - **OpenAI CLIP collection50**
  - **PyTorch Taskonomy** (visualpriors) project
  - **Detectron2** model zoo
  - **Harvard Vision Sciences Laboratory's Open-IPCL** project

- Variations across models in terms of architecture and training objectives



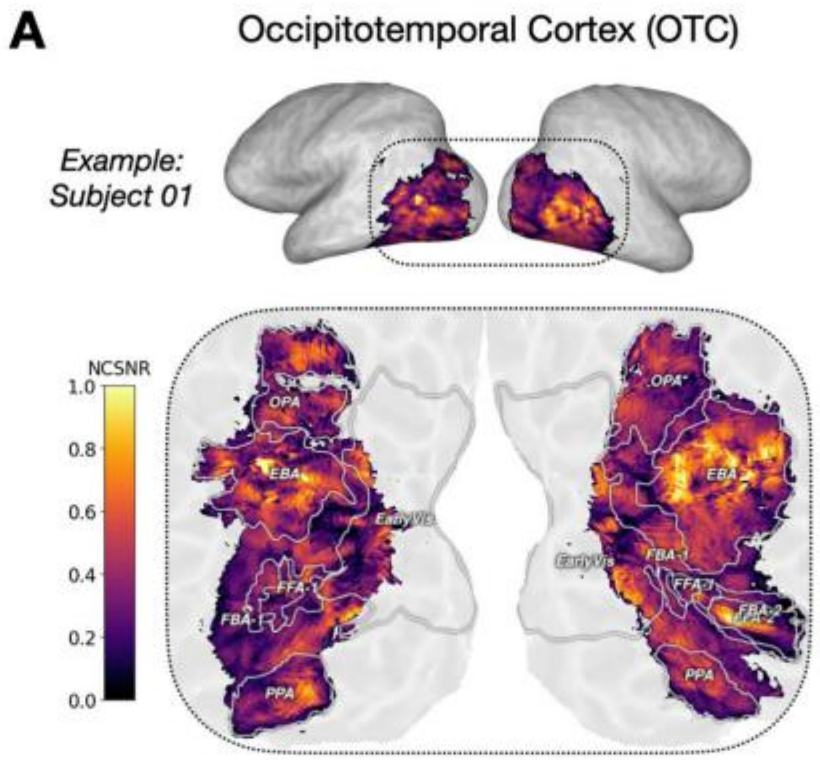
## Human fMRI Data

- Natural Scenes Dataset
  - **70k** visual stimuli
  - Images from **COCO dataset**
  - Resolution: **7T** field strenght, **1.6-s TR**, **1.8mm$^3$** voxel size
  - **4** subjects (01, 02, 05 and 07)
  - **1000 stimuli overlap** between subjects



## Voxel Selection Procedure

- For high **SNR**
  - Used **NCSNR** (noise-ceiling SNR) to select **reliable voxels**
  - Threshold used = 0.2
- For **ROI**
  - **Occipito-temporal cortex** (OTC)
  - **Broad Mask:** Selected from the **"nsdgeneral" ROI** (visual system).
  - **Refined Selection:** Kept voxels from **mid-to-high ventral & lateral ROIs**.
  - **Category-Selective ROIs:** Included voxels from **11 face, body, word, scene ROIs**

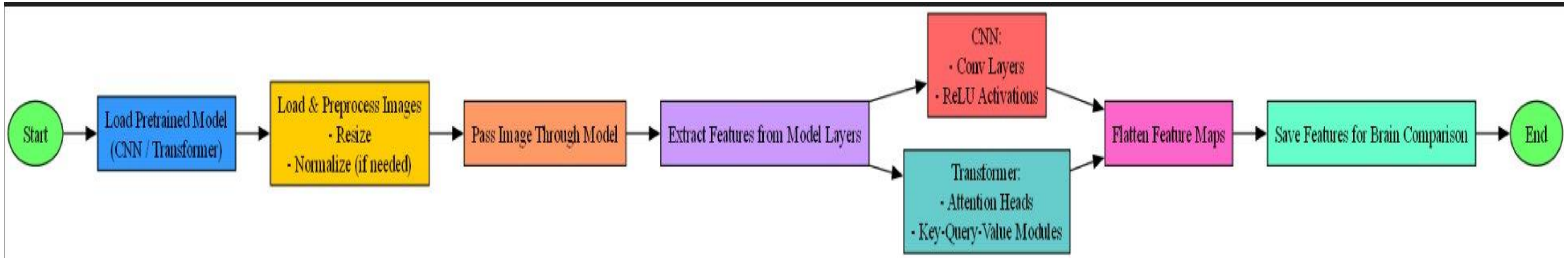


## Noise Ceiling

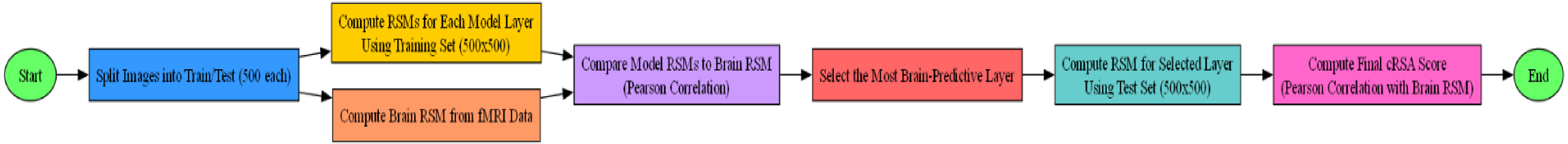
- Maximum possible achievable performance given the noise in data
- In current context – implies within-subject RSMs where variability across trials is impacted by noise
- Novel method – GSN – generative modelling of signal and noise
- Estimates multivariate gaussians over an ROI assuming that observed data contains additive nature of noise samples
- Post-hoc scaling of signal distribution to match empirically observed reliability of RSMs
- Noise ceiling estimation by correlating noise-less RSM (gen.) with estimated RSM (using sig.-noise estimation)

# Feature Mapping methods

- All probe images are tensorized via the "test-time" transformation of the given model , for untrained models this is skipped , for no available transformation , they reconstructed the transformation required.
- Feature extraction: feature maps extracted from CNN layers before and after activation , from Transformers , each attention head separately and each KQV modules inside the attention heads.
- Finally for each model's each layer they have a feature matrix of dimension , num_images*num_features, (flattened from the original feature map)



# cRSA


Start
Split Images into Train/Test (500 each)
Compute RSMs for Each Model Layer Using Training Set (500x500)
Compute Brain RSM from fMRI Data
Compare Model RSMs to Brain RSM (Pearson Correlation)
Select the Most Brain-Predictive Layer
Compute RSM for Selected Layer Using Test Set (500x500)
Compute Final cRSA Score (Pearson Correlation with Brain RSM)
End

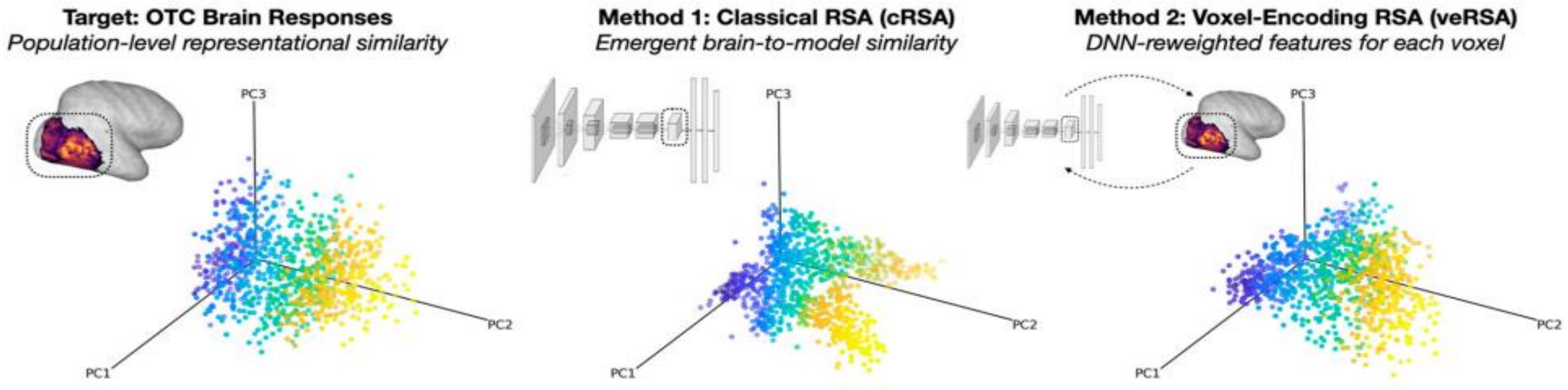
# veRSA

- The pipeline is like cRSA pipeline, except for the addition of the encoding procedure
- Steps involve->
- Dimensionality reduction using Sparse random projection(JL lemma) , Then training the encoding model for each voxel using L2/ridge regressor.
- With these weights we predict responses for each voxels and use these to get a predicted RSM, select the most predictive layer , then run the test set to get the RSA of this layer with the brain data,

# Results

# Comparision metric

Comparision of Population geometry of fMRI data to Predicted Population geometry(from the model being tested) done by Representational similarity analysis using 2 different methods(cRSA and veRSA)

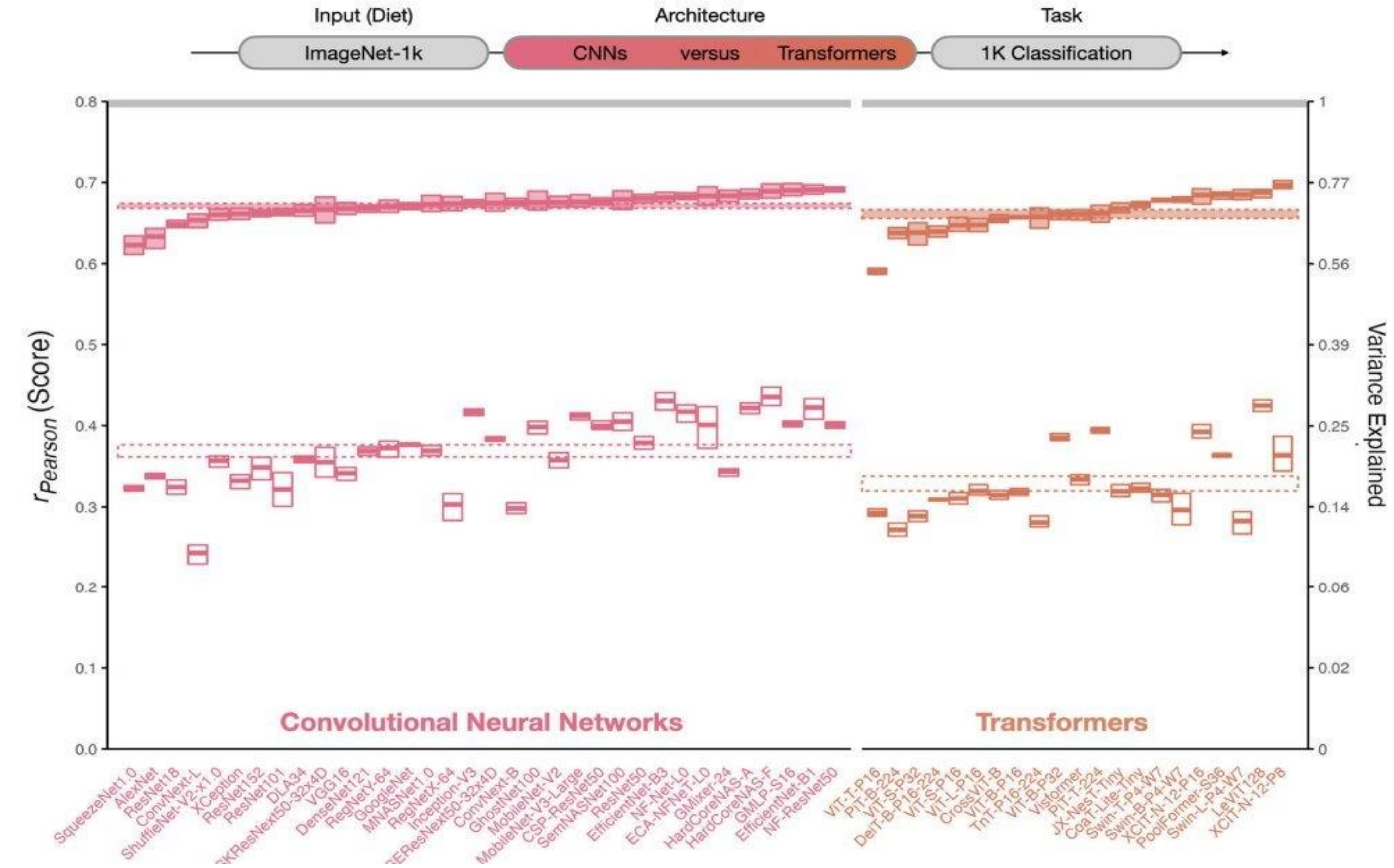


# Architecture Comparison

Instead of focusing on low-level differences like number of parameters, layers, width of layer, batch size etc.,

The architectures chosen for comparison are different in a meso-scale architectural motif (which means a medium-level design feature that fundamentally affects how the model processes information) in this case being ***<u>Convolutional Bias</u>***

CNNS have this bias while Pure Vision transformers donot have this bias , so these 2 architectures are chosen for the ablation of Architectural comparison

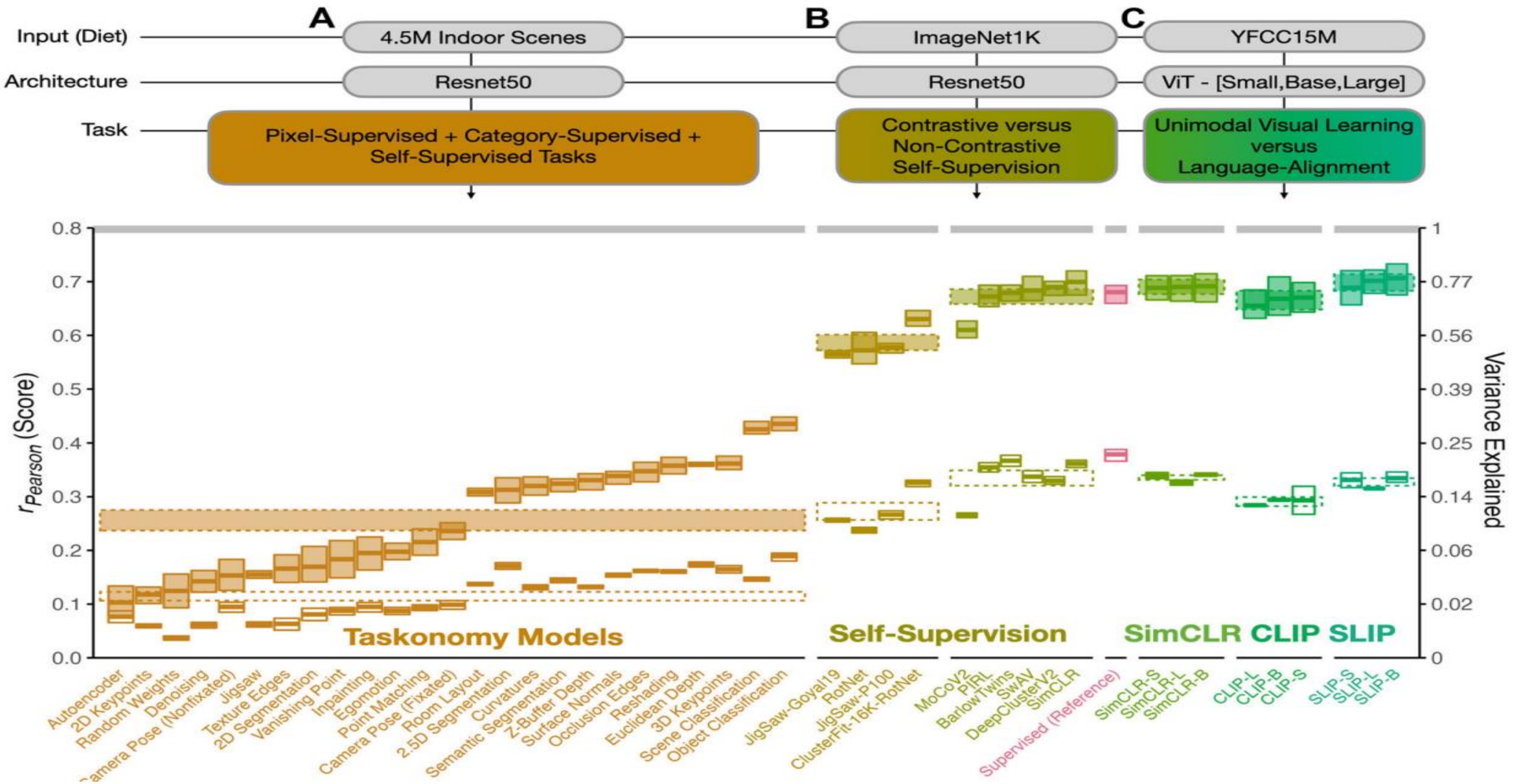


Input (Diet)
ImageNet-1k
Architecture
CNNs
versus
Transformers
Task
1K Classification
$r_{Pearson}$ (Score)
Variance Explained
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
0.77
0.56
0.39
0.25
0.14
0.06
0.02
0
Convolutional Neural Networks
Transformers
SqueezeNet1.0
AlexNet
ResNet18
ShuffleNet-V2-x1.0
ConvNext-L
XCeption
ResNet152
ResNet101
DLA34
SKResNext50-32x4D
VGG16
DenseNet121
RegNetY-64
GoogleNet
MNASNet1.0
RegNetX-64
Inception-V3
SEResNext50-32x4D
ConvNext-B
GhostNet100
MobileNet-V2
MobileNet-V3-Large
CSP-ResNet50
SemNASNet100
ResNet50
EfficientNet-B3
NF-Net-L0
ECA-NFNeT-L0
GMixer-24
HardCoreNAS-A
HardCoreNAS-F
GMLP-S16
EfficientNet-B1
NF-ResNet50
ViT-T-P16
PiT-B-224
ViT-S-P32
DeIT-B-P16-224
ViT-S-P16
ViT-L-P16
CrossViT-B
ViT-B-P16
TnT-P16-224
ViT-B-P32
Visformer
PiT-T-224
JX-NesT-Tiny
CoaT-Lite-Tiny
Swin-T-P4-W7
Swin-B-P4-W7
XCIT-N-12-P16
PoolFormer-S36
Swin-L-P4-W7
LeViT128
XCIT-N-12-P8

- On average , Both CNN and Transformer models predicted responses equally well with Transformers doing slightly worse
- Leading to the hypothesis that CNNS might be introducing inductive biases relating to the more brain-aligned representations , but this can't be taken as a claim since the predictions range were largely overlapping
- Found the cRSA score to be vastly lower than veRSA scores leading to the hypothesis that the veRSA was remapping the representations of both the models to similar sub-spaces more aligned with the brain's representations after reweighting.

# Task Comparision

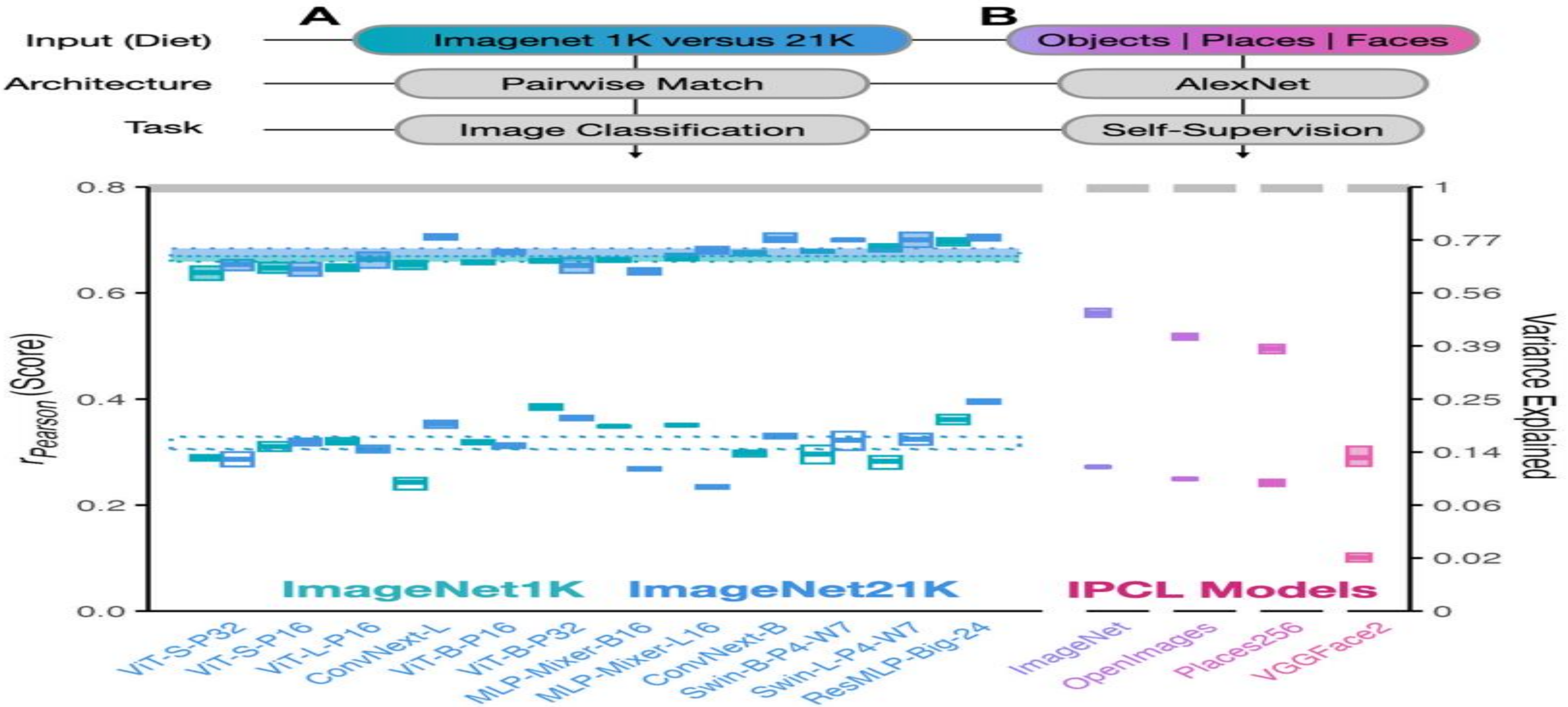
Task variation done across Taskonomy models, Self-Supervised models,language alignment SLIP models



- The taskonomy models performed the worst , with the lowest being autoencoders and the highest being Object detectors , the Object detector performed worse than an Image-net-1k Resnet 50 despite being trained on a larger dataset, conclusively saying that the diversity of the dataset matters (Taskonomy has only about 100 of the 1000 classes of images compared to imagenet)
- The contrastive SSL models performed better than non-contrastive ones, even performing as well as full-supervised , meaning brain processes visual information like the contrastive models (i.e, Learning invariances in similar images)
- The Pure language aligned models (CLIP) performed worse than SLIP(hybrid) and SLIMCLR(pure self-supervision), leading the authors to believe that the good performance of OPEN-AI's CLIP was due to the large undisclosed dataset of 400M images rather than influence of language alignment.

# Input Comparision

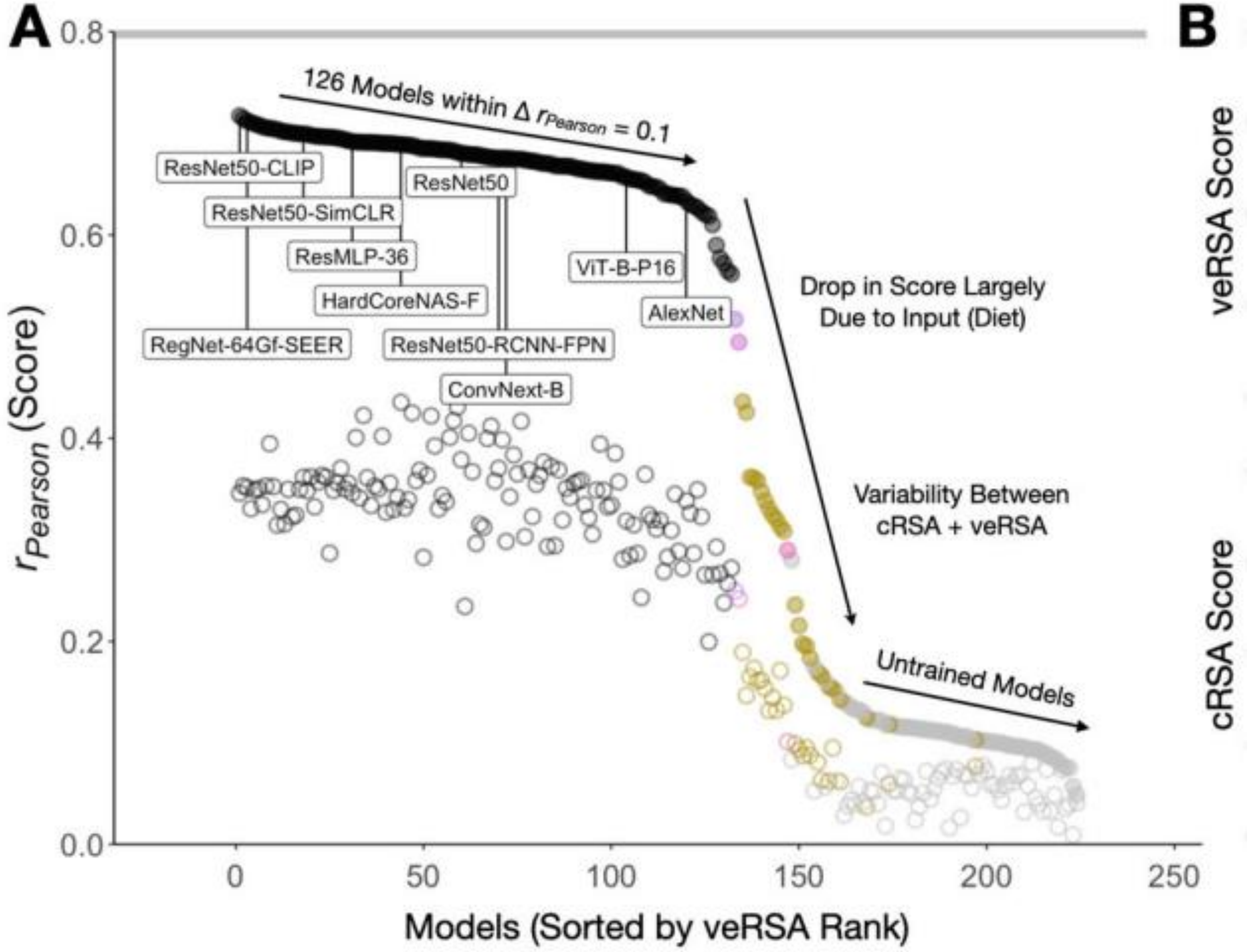
Input variation done over Imagenet1k v/s 21k , and on IPCL models trained on Imagenet, OpenImages, Places365 and VGGFace2



- No major differences in cRsa and veRsa scores of Imagenet 1k and imagenet 21k , suggesting that training on more images isnt always better, and that the improved apparent diversity didn’t translate into better representations.
- IPCL model ranked from best to worst -> Imagenet,OpenImages(object focused),Places365(scene focused),and VGGFace2(face-focused),
- Despite Imagenet having lesser images than the other datasets it outperformed the other 3 , hinting that the latent dataset diversity of Imagenet is higher than these.
- Significant differences in performance suggest that visual diet plays a major role in brain-predictive power of the model.

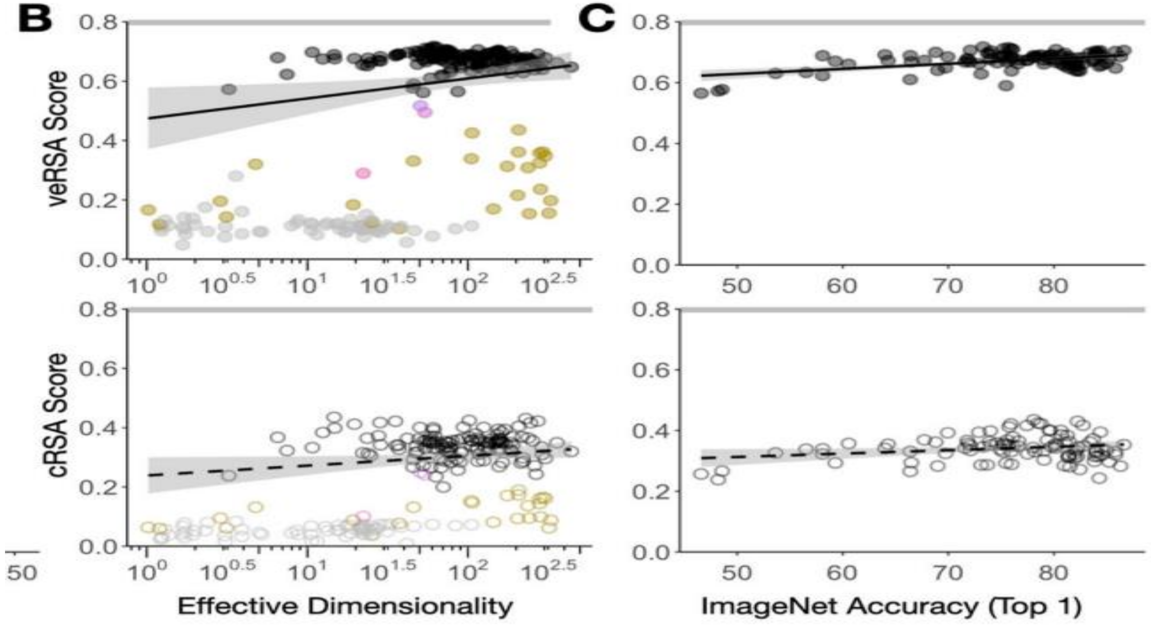
# Impact of training, overall comparison of models

- Comparision of random initialisation vs pre-trained models for brain predictivity.



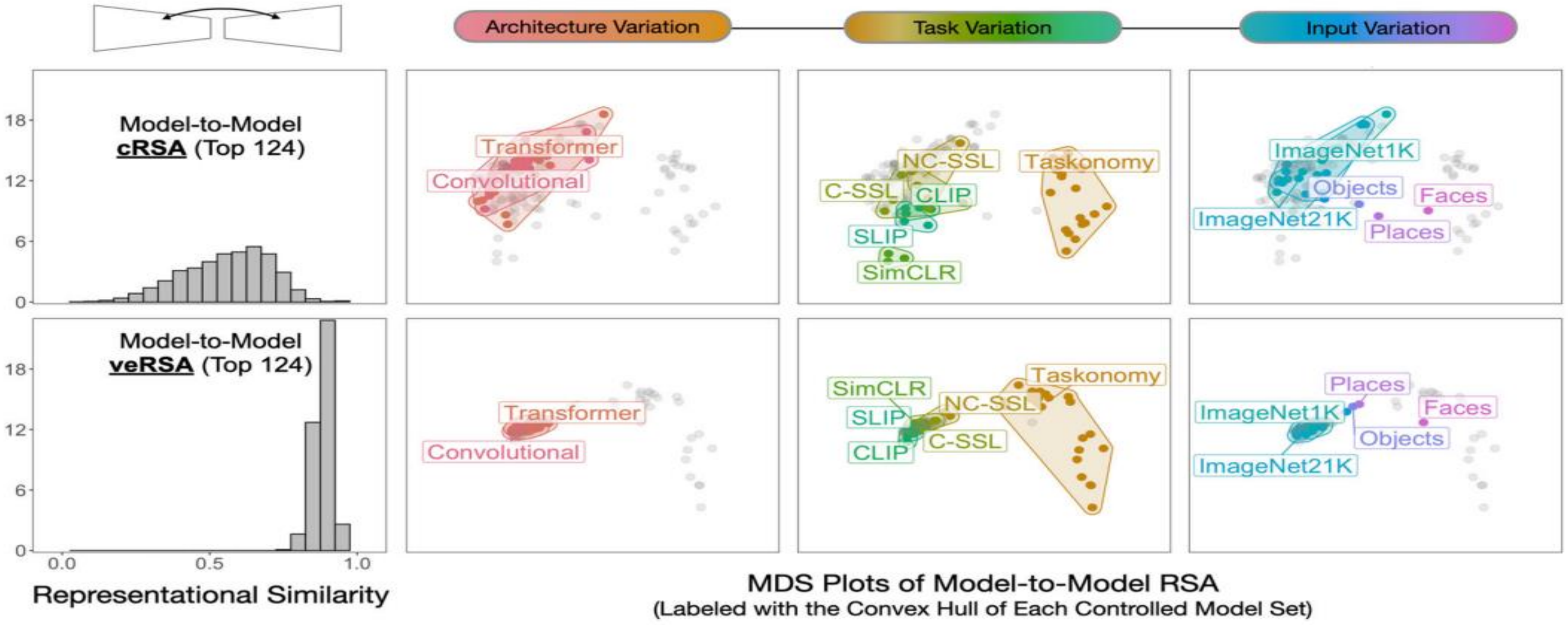
- The impact of training is shown as the untrained model perform worse than trained model
- Overall variation is that 126 models out of 224 performed relatively well , then followed by the taxkonomy models (less diverse dataset) and then the untrained models
- Next we diversify on effective dimensionality , classification accuracy, and number of trainable parameters

- Variation of ED in trained models and random models vs  brain predictability showed no correlation when compared separately ,concluding that ED is not an indicator of brain-predicting power
- Little to no correlation found between classification accuracy and predictability concluding that brain predictavity isnt indicated by the fine-tuned weights of the top-1 imagenet model.
- No consistent behaviour found for number of trainable parameter variation having an effect on cRsa and veRsa scores.



# Model to model comparision

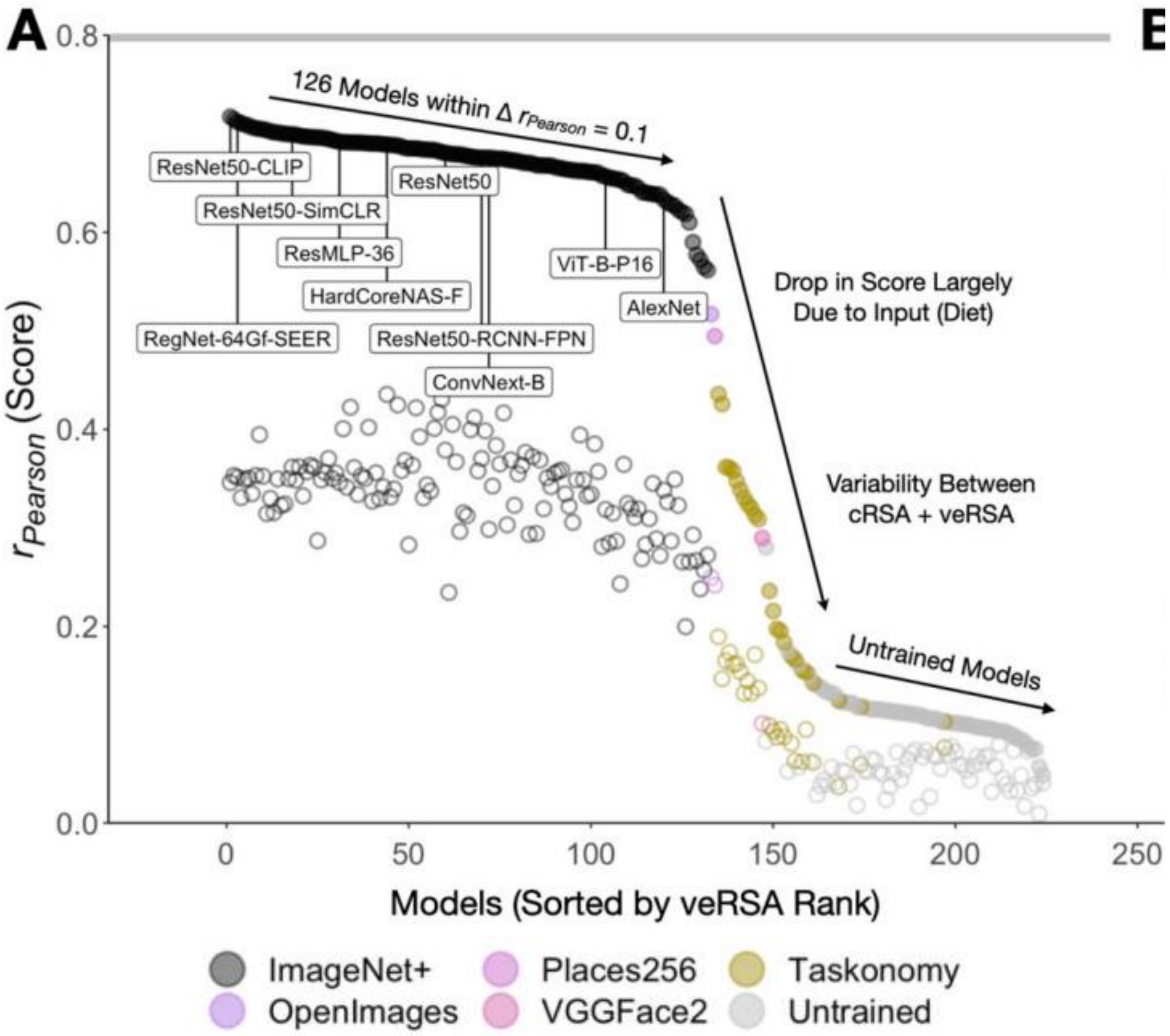
To understand that whether differences in architecture,task,diet actually lead to differences in representation , a model similarity analysis was run , done using cRSM and veRSM



- Substantially variant and diverse representations in the cRSA models,
- The variation collapsed for veRSA models , suggesting that reweighting plays a crucial role for models to converge into a shared-brain like representation.
- Stating that the difference between cRSA and veRSA is the most significant predictor of brain responses when compared to all the inductive biases covered so far.

# Discussion

# Importance of Visual Experience



# Importance of Visual Experience

Two key observations:

- Untrained models (without any visual experience) not able to capture later-stage representations
- Impoverished diets (data) are quite bad for models

However, few issues:

- No metric or measure for knowing input data richness – available for measuring image similarity in latent space (too late!)
- Use this to perform "semantic deduplication" using embedding space of models like CLIP
- Another issue is of non-uniform data-augmentation and set of hyperparameters

**Model-to-brain linking**

- Discussion about possible flaws of metrics , if veRSA is better, then the representations learned by CNN vs ViT might not matter much but if cRSA is better than these models do learn different representation , it's just that the metric can't capture it properly,
- Provided direction regarding future metrics aligning with Sparsity constraints , mapping from multiple layers (to display heirarchy) , one-to-one mapping
- Critiqued that it could be the Dataset which is flawed , NSD might not provide the images that could bring out the representational differences between these fine-tuned models
- Usage of artificial stimulus (generated images) to differentiate the models' representations from each other

# Novelty

- As mentioned in the paper:
  - Large model scopus
  - Diverse datasets
  - Different tasks and training paradigms
  - Model-to-Model comparision
  - Emperical derivation of a set of shared inductive biases based on which they perform statistical grouping of models
- Previously, only a single model feature was explored with regard to model-to-brain alignment and tested on conventional benchmarks
- Previously, more attention given to building a go-to model for brain

# Conclusion

## Limitations

- Limited exploration in terms of mapping/alignment
- Not explored heirarchical mapping of model and brain – different stages of model to different brain regions
- Much less focus on category-selective regions
- Focus on later stages of visual processing – little to no exploration about early stages of models as well as visual stream
- Variability in model-dataset pairs – not all models tested across similar datasets – groupwise done

## Future Directions

- Implement and explore heirarchical mapping of models to brains
- Better analysis of "visual diet" for training the model

What we thought can be added:

- Multimodal datasets for long range dependency, contextual augmentation and how they impact visual stream