Bioinformatics institute

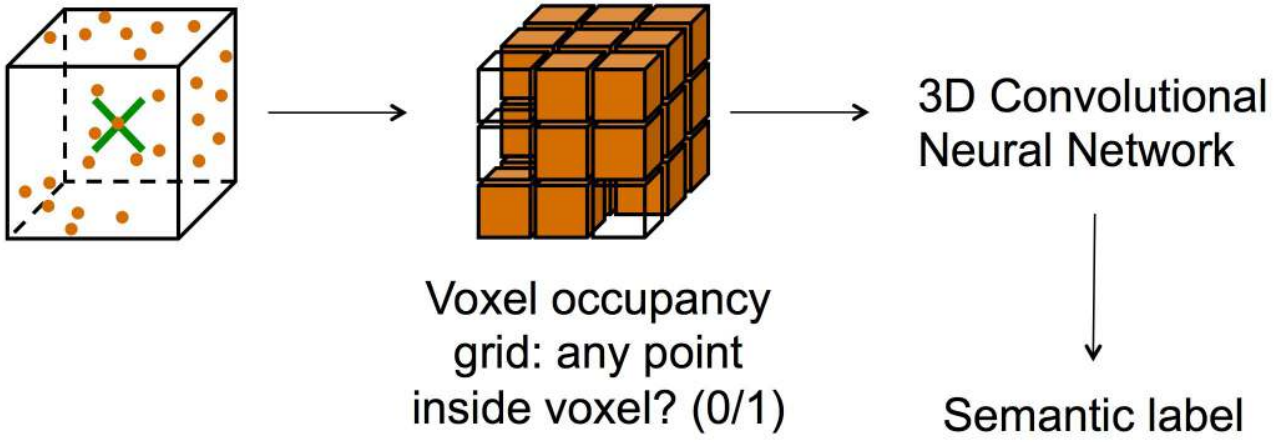
# Deep learning in protein anomalies

**Supervisor:**
Andrey
Afanasiev
iBinom

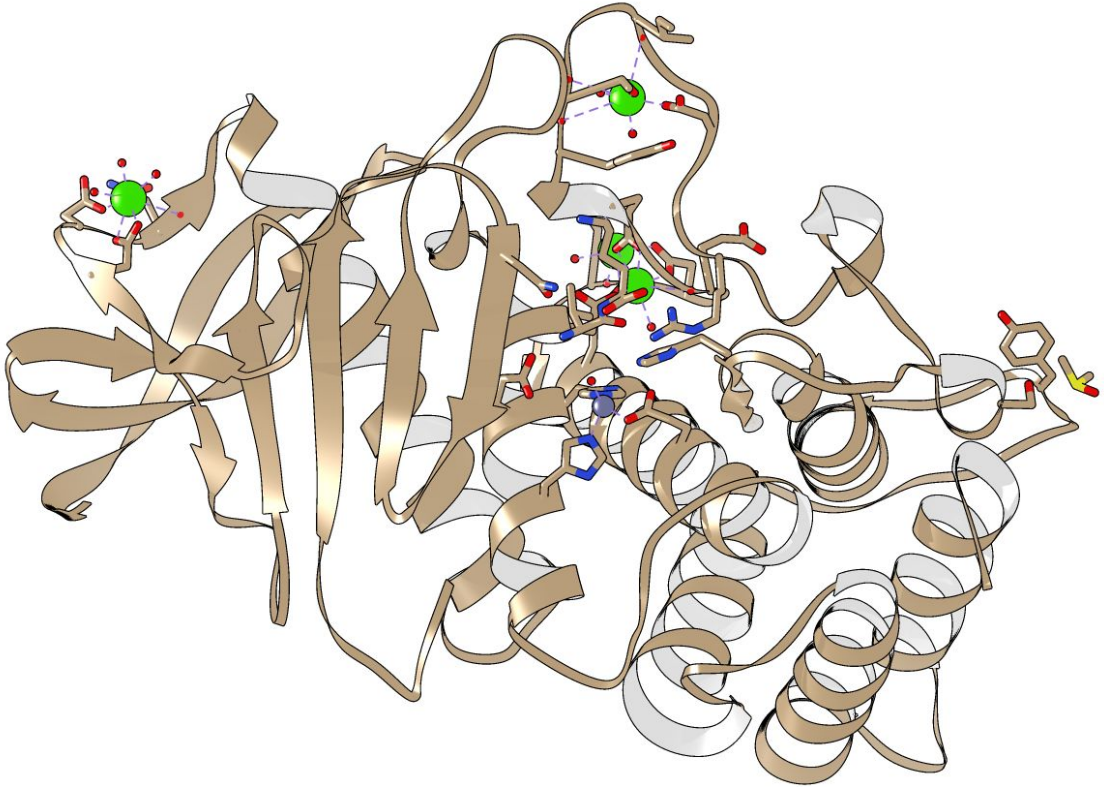
**Student:**
Ivan Reveguk

# Initial idea is...

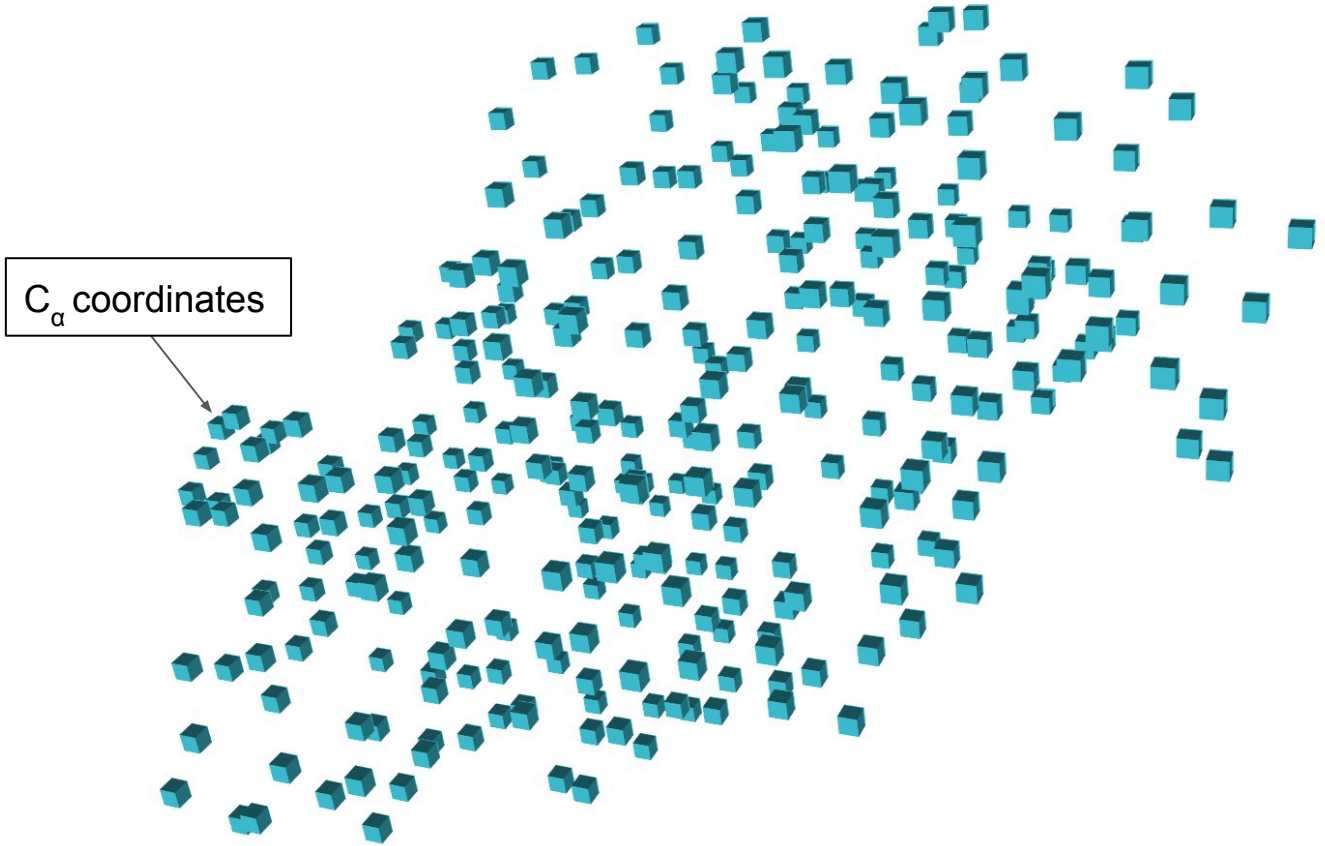
**Using 3D convolutional neural networks train a model to predict a single amino acid substitution's impact on protein function**



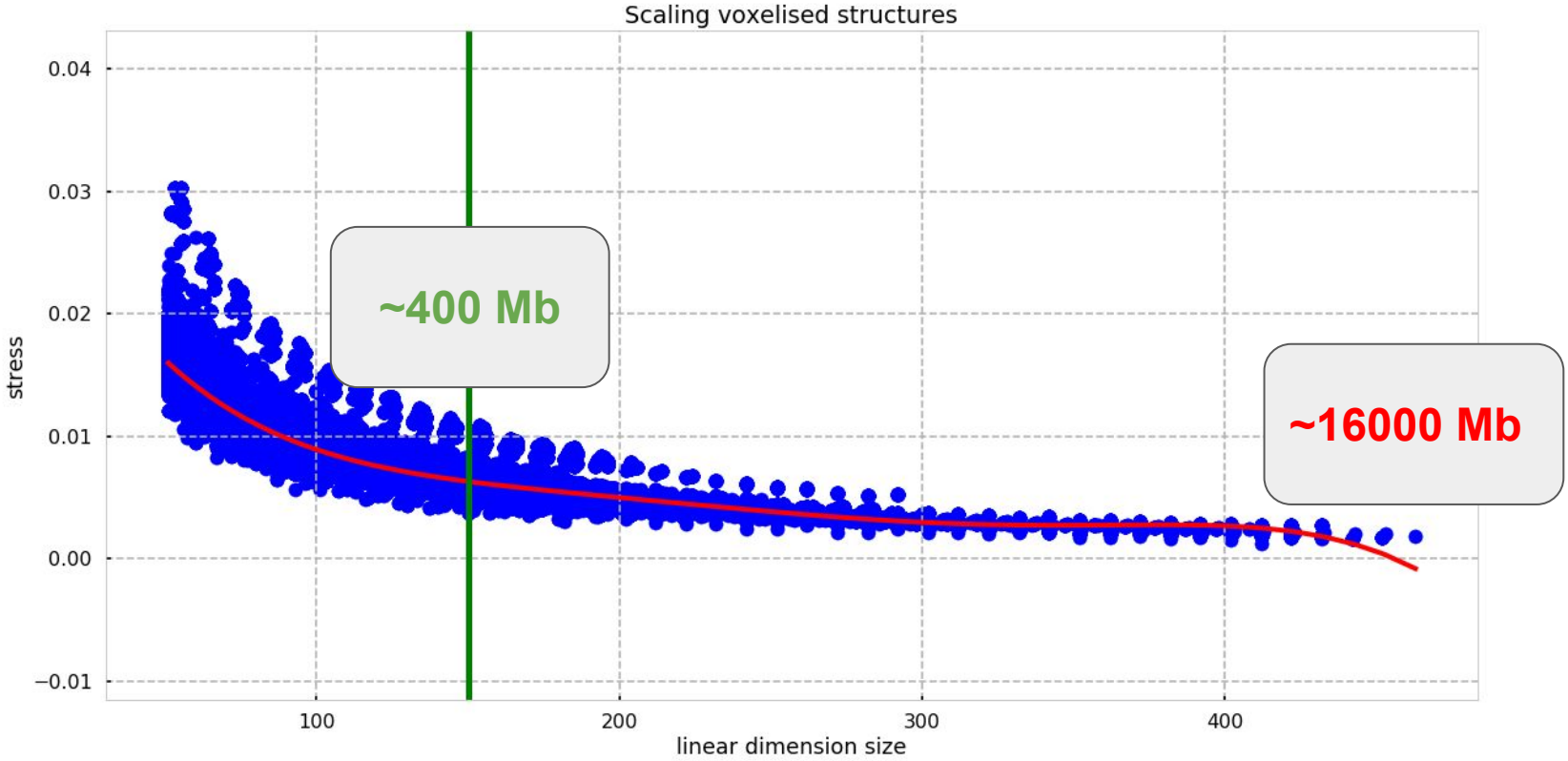
# Structure representation



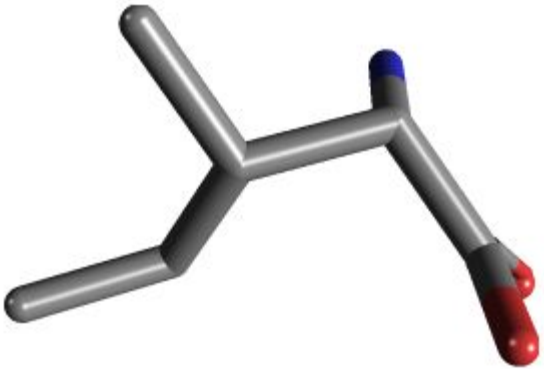
# Voxel-based protein models

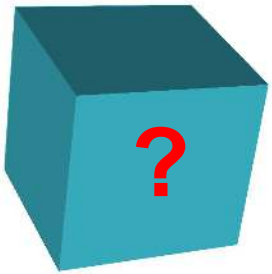


# Memory efficiency problem



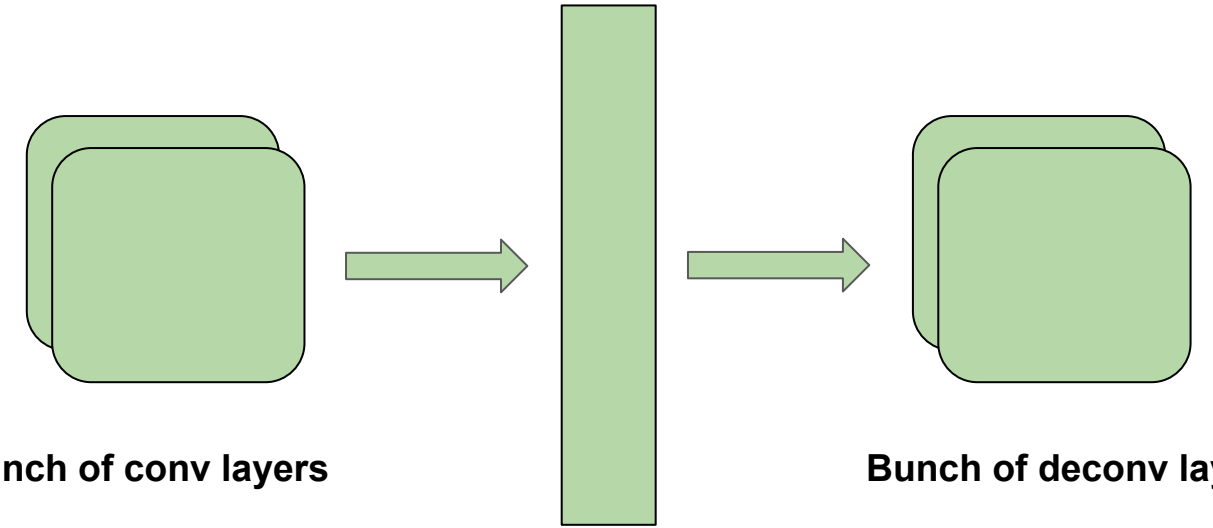
# Amino acid representation problem





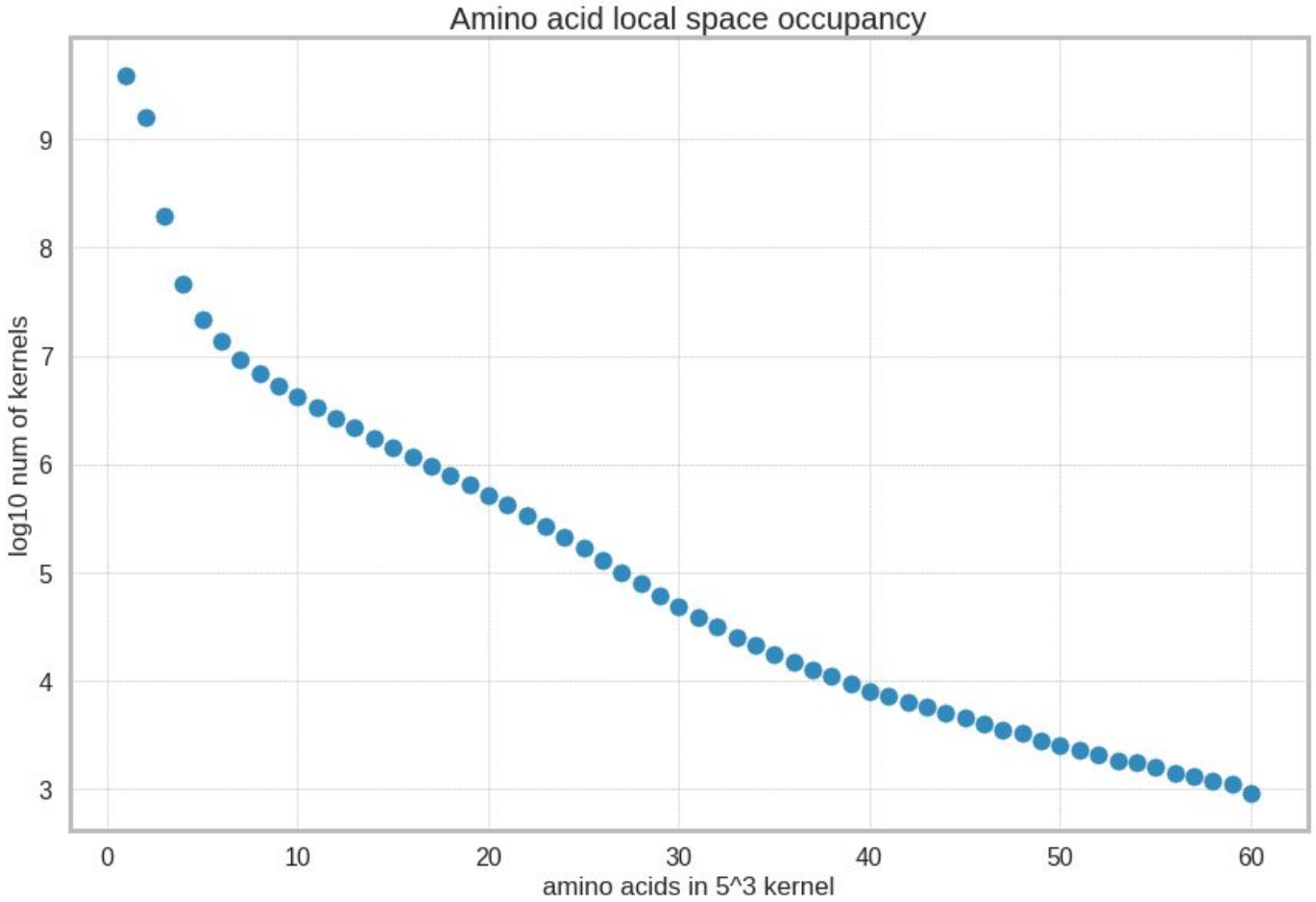

Volume
Hydrophobicity
Charge
...

# Testing on autoencoder

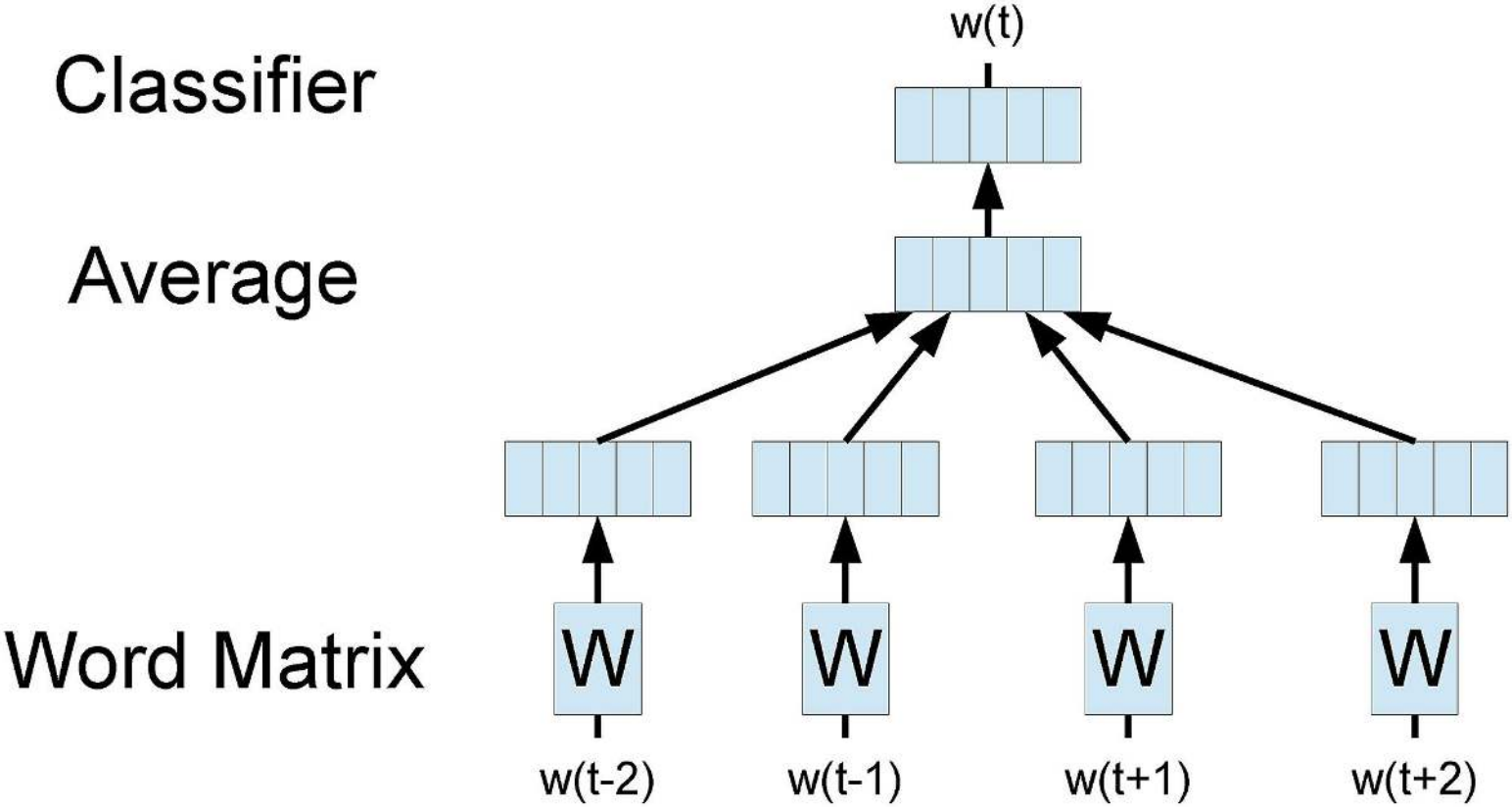



**Poor results. Why?**

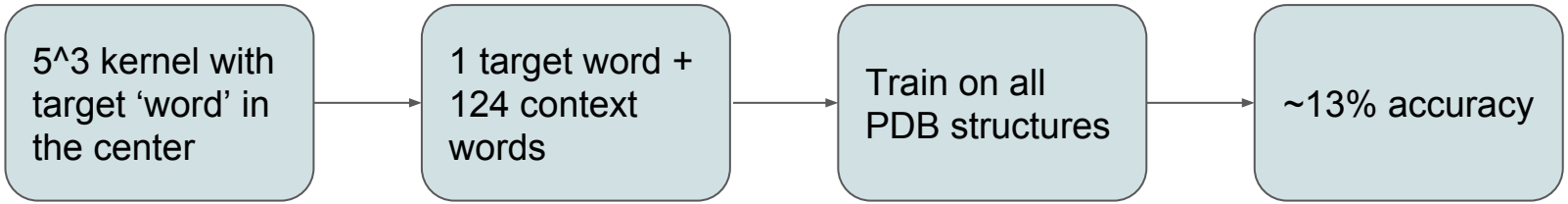
# Empty space problem

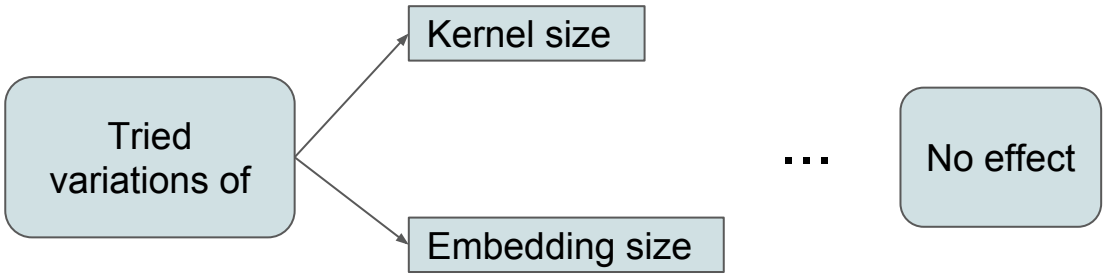

Amino acid local space occupancy
log10 num of kernels
9
8
7
6
5
4
3
0
10
20
30
40
50
60
amino acids in 5^3 kernel

# Vector embeddings
# CBOW model



# Vector embeddings



# Current results

- **Personal**

  A lot of skills gained; got my hands real dirty:

  Deep learning from zero

  A lot of data wrangling and literature reading during side quests

- **Objective**

  Addressed some of the most important technical issues in representation of protein molecules with preserved spatial information:

  - Empty space problem makes it really hard to use conventional channels in CNN architecture or train vector embeddings in 3D
    - Analyzed space occupancy for all PDB proteins
    - Trained CAE
    - Trained vector embeddings
  - Size of a model can be reduced by scaling
    - Manually chosen optimal size 150x150x150
    - 40x improved  memory efficiency

# Future directions

- Find better ways of representation, while preserving spatial structure:
    - Dimensionality reduction (3D -> 2D) (diffusion maps)
    - Train embeddings in 2D
- Proceed with 3D CNNs