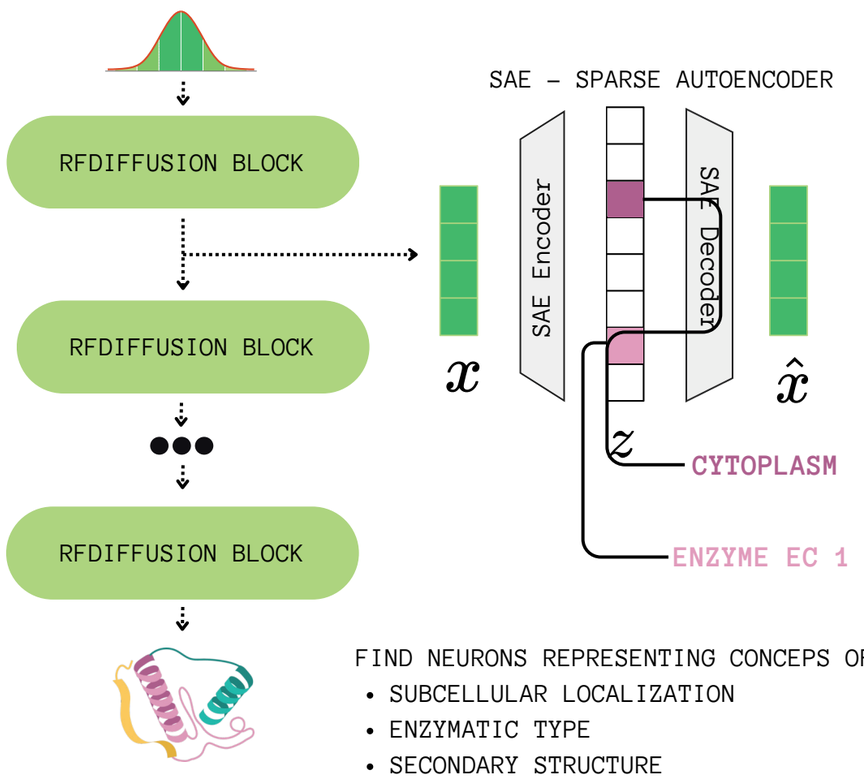
RFSAE INTERPRETABLE PROTEIN BACKBONE DESIGN USING **RF**DIFFUSION AND **S**PARSE **A**UTO**E**NCODERS

Wojciech Zarzecki, Paulina Szymczak, Ewa Szczurek, Kamil Deja

01 DISENTANGLING FEATURES WITH SAE



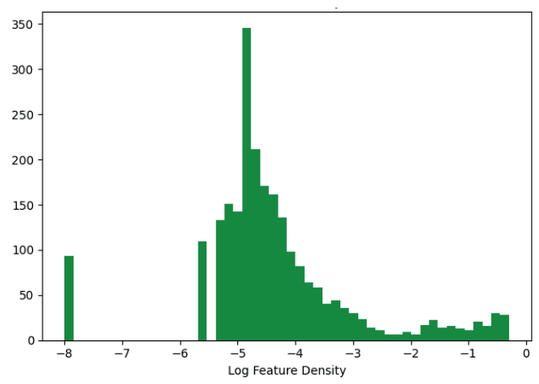
FIND NEURONS REPRESENTING CONCEPS OF:

- SUBCELLULAR LOCALIZATION
- ENZYMATIC TYPE
- SECONDARY STRUCTURE

02 ABLATION OF BLOCKS

WHICH BLOCK STORES KNOWLEDGE ABOUT THIS CONCEPTS?

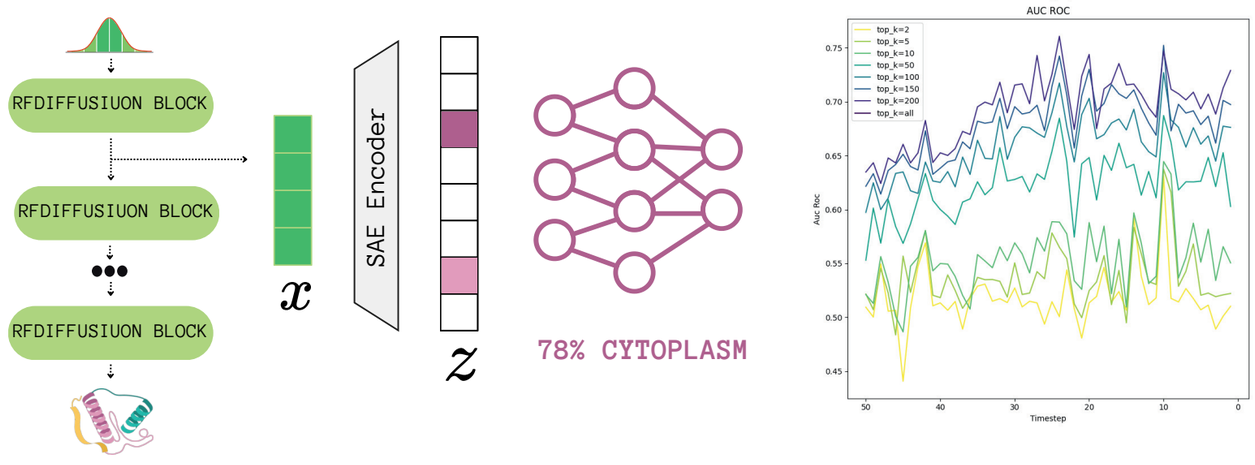
03 ENFORCING SPARCITY



$$\mathbf{z} = \mathrm{TopK}\left(W_{\mathrm{enc}}\left(\mathbf{x} - \mathbf{b}_{\mathrm{pre}}\right)\right)$$

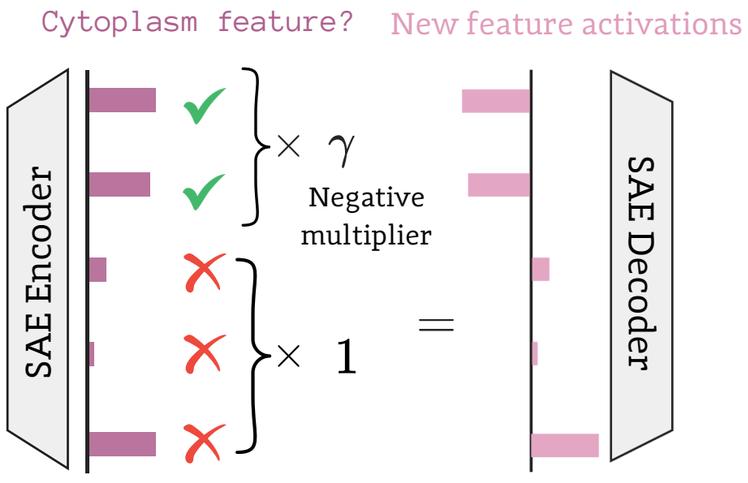
HOW TO MAKE LATENT SPACE SPARSE?

04 PROBING MODELS



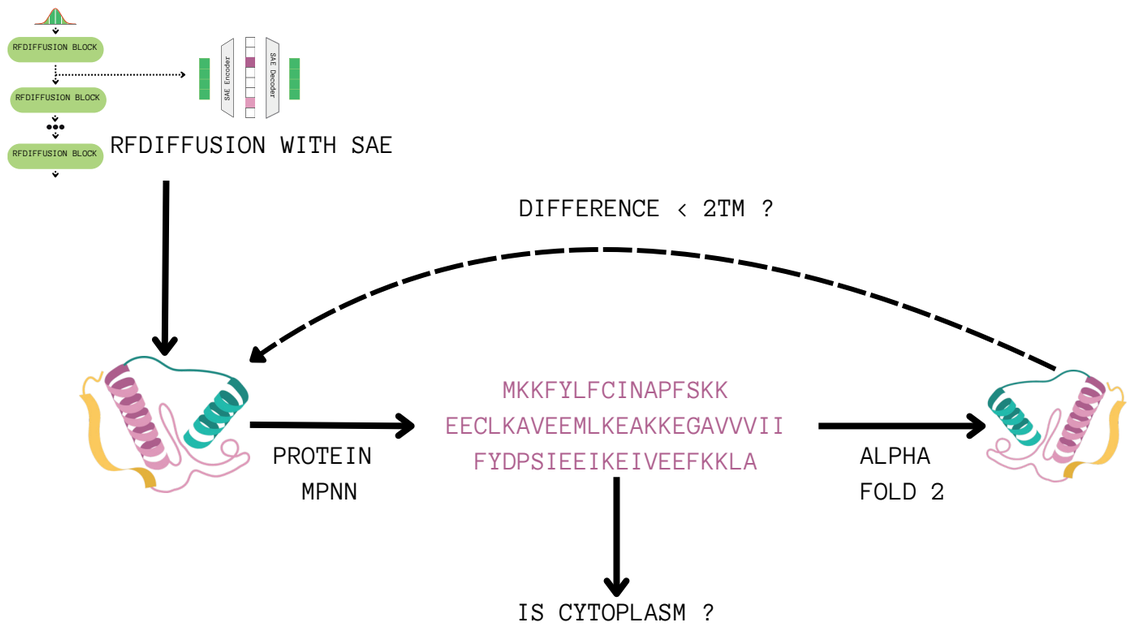
NEURONS REPRESENTING CONCEPS = IMPORTANT FEATURE OF PROBING MODELS

05 INTERVENTION AT INFERENCE



BLOCK OR REFINFORECE INDENTIFIED NEURONS

06 EVALUATION PIPELINE



UNIVERSITY OF WARSAW
Warsaw University of Technology
HELMHOLTZ MUNICH