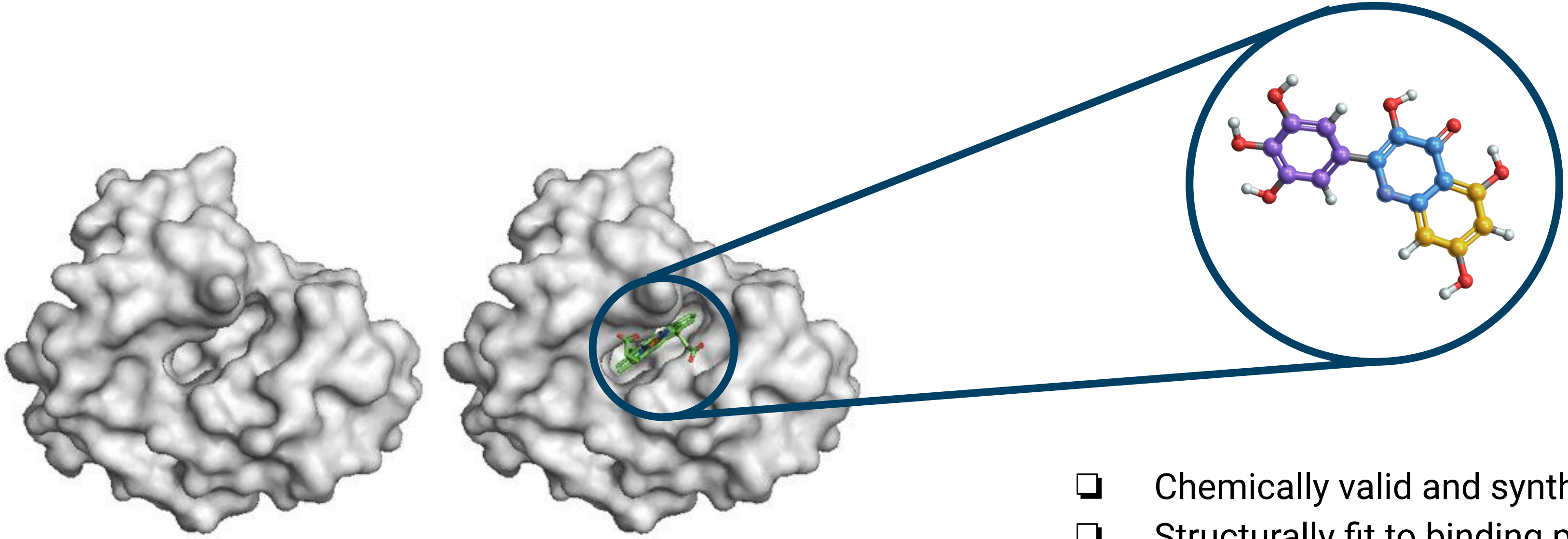
ImageToMolecule

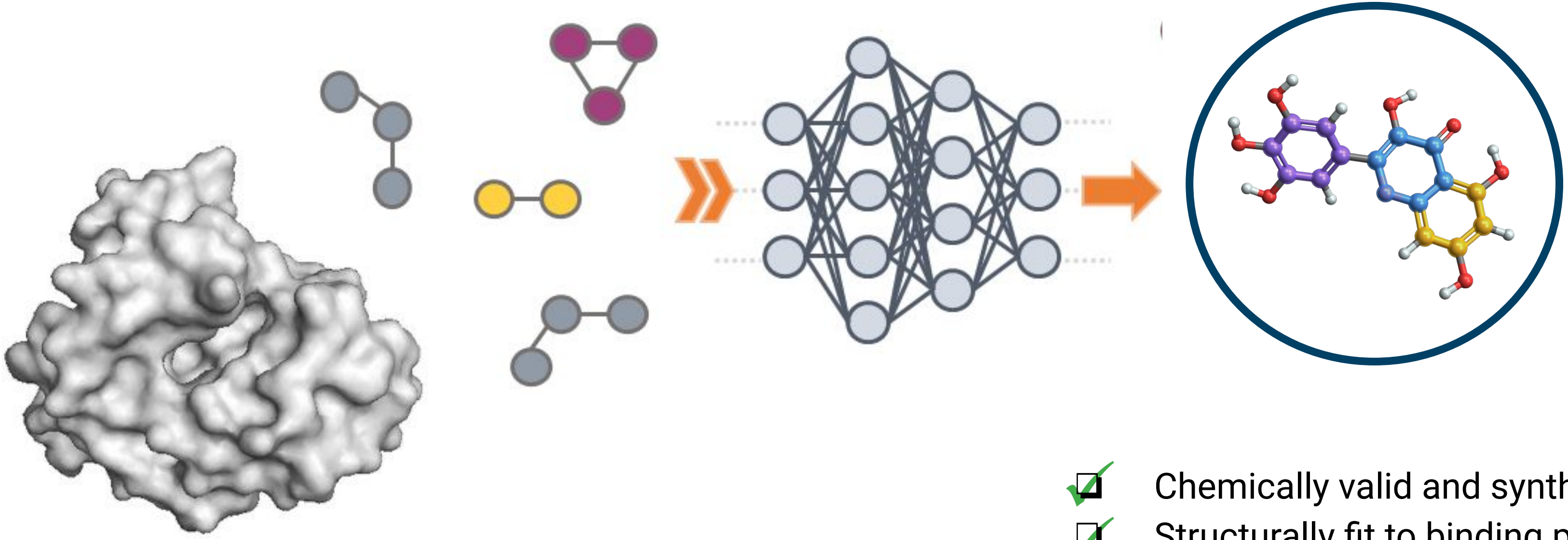
Learning Protein Localization Images for Biologically-Specific Molecular Design

Small molecule drug discovery is hard



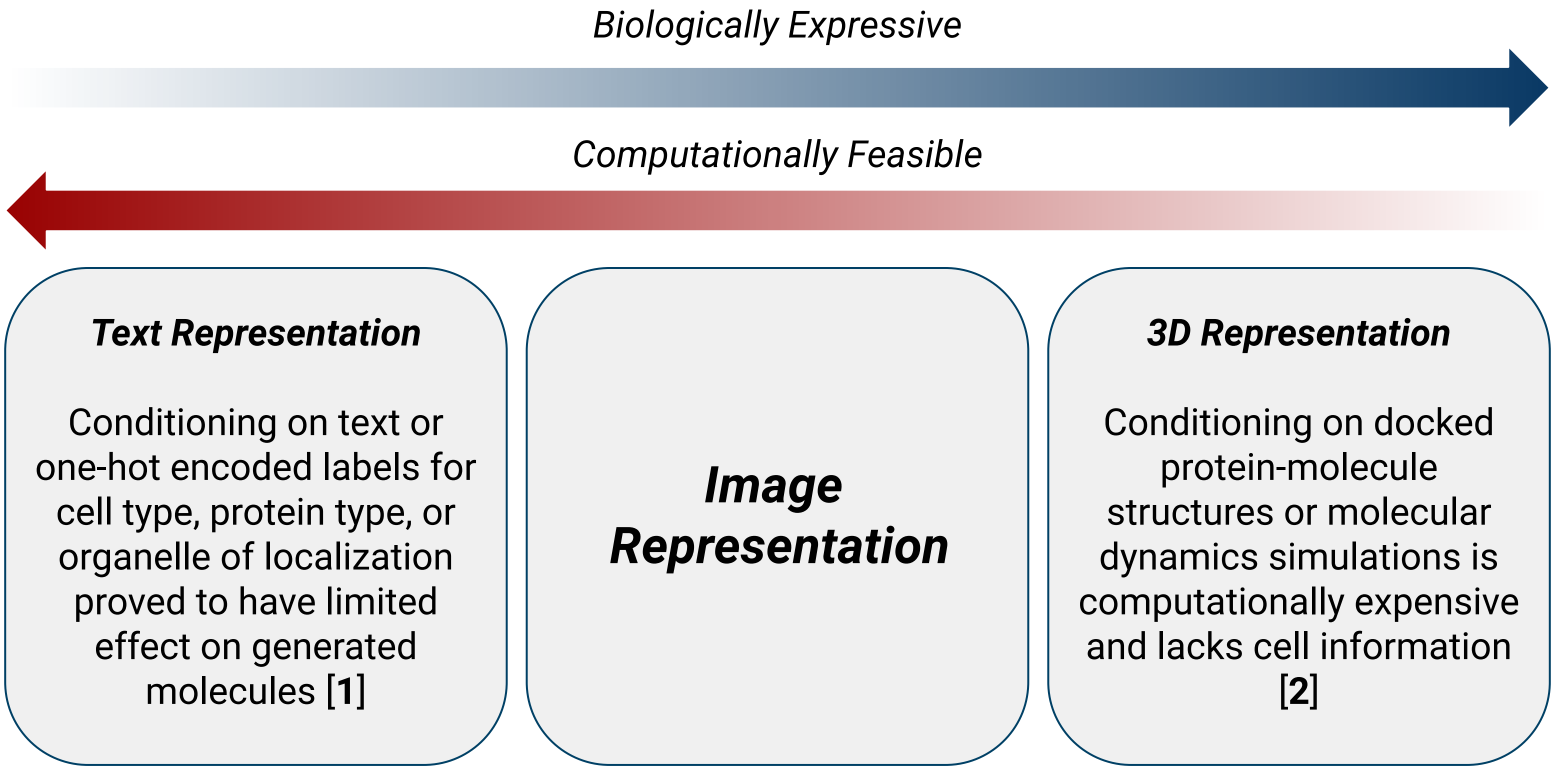
- ❑ Chemically valid and synthesizable
- ❑ Structurally fit to binding pocket
- ❑ Active in target cell & organelle

Generative models help, to some extent



- ☑ Chemically valid and synthesizable
- ☑ Structurally fit to binding pocket
- □ Active in target cell & organelle

Trade-offs when integrating biological context



Biologically Expressive

Computationally Feasible

Text Representation

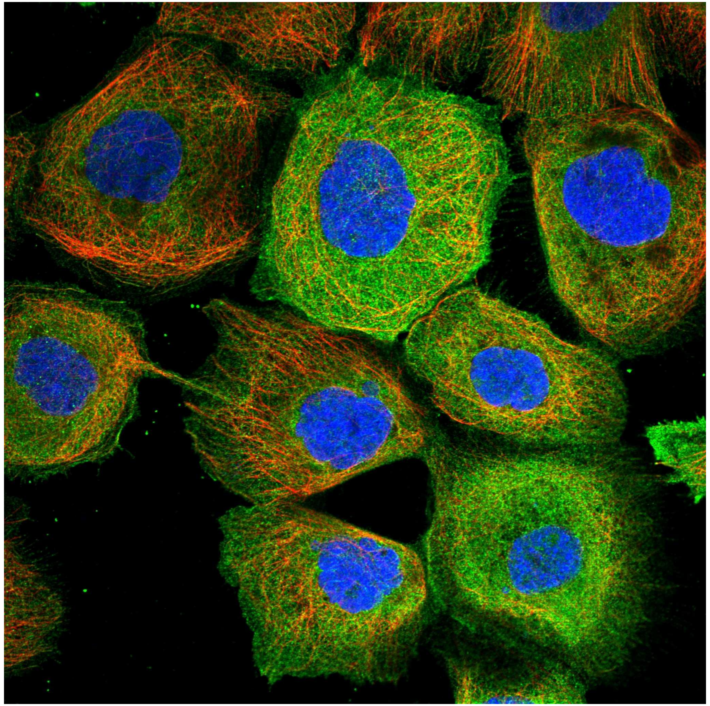
Conditioning on text or one-hot encoded labels for cell type, protein type, or organelle of localization proved to have limited effect on generated molecules [1]

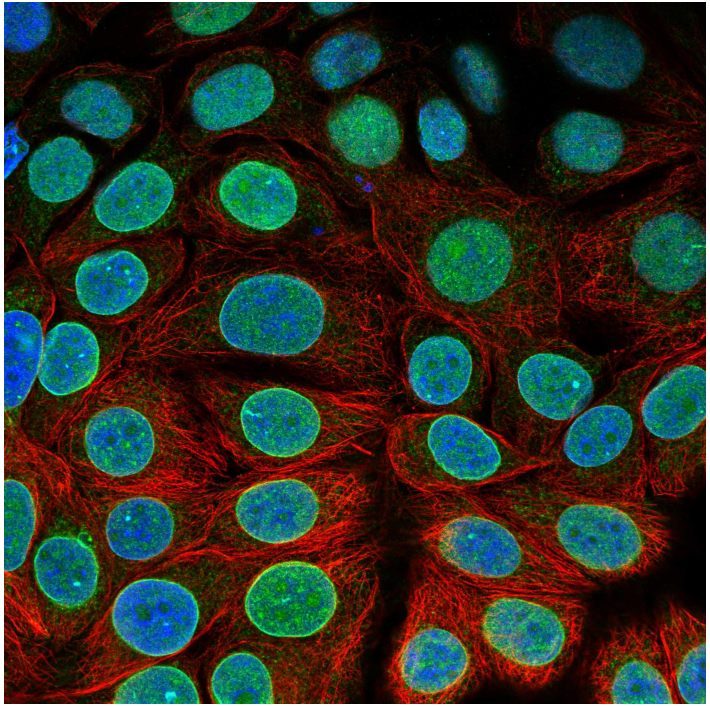
Image Representation

3D Representation

Conditioning on docked protein-molecule structures or molecular dynamics simulations is computationally expensive and lacks cell information [2]

Are protein localization images the answer?

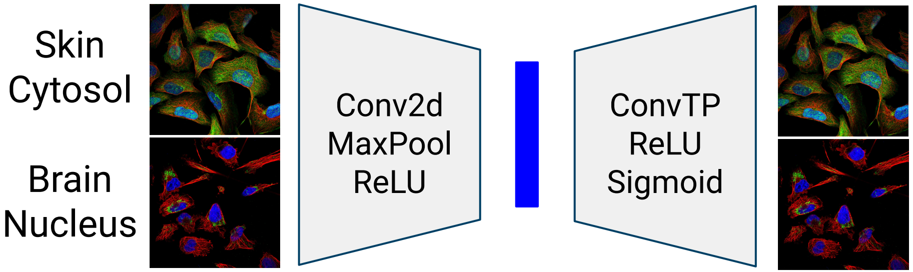




Algorithm overview

Convolutional Autoencoder [3]

Loss: Reconstruction + Contrastive [4]



Skin Cytosol

Brain Nucleus

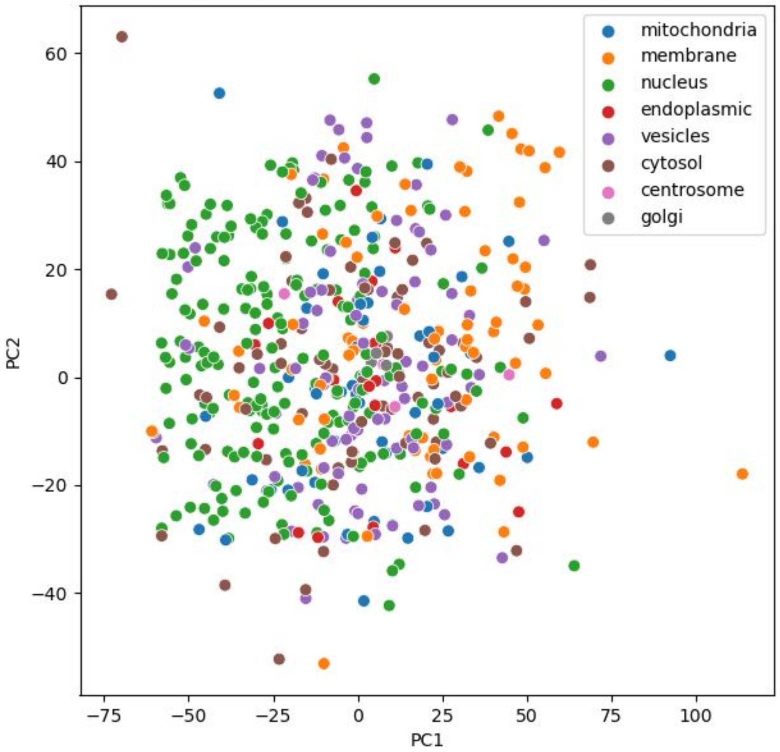
Conv2d
MaxPool
ReLU

ConvTP
ReLU
Sigmoid

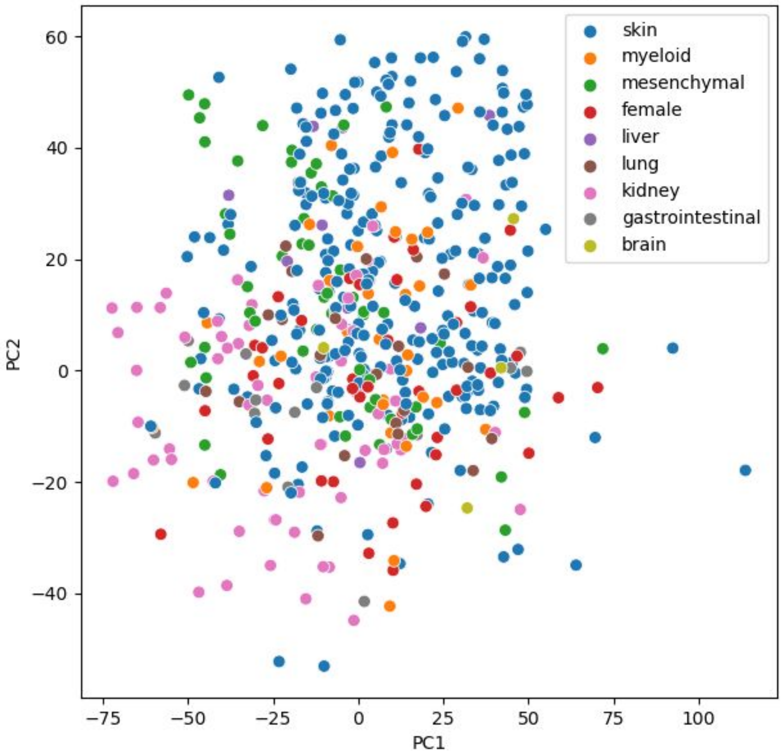
$$L = MSE(\hat{y_1}, y_1) + MSE(\hat{y_2}, y_2) + \\ l_{org} * d^2 + (1 - l_{org}) * max(0, 0.1 - d)^2 + \\ l_{ct} * d^2 + (1 - l_{ct}) * max(0, 0.1 - d)^2$$

l_{org}	1 if same organelle, 0 otherwise
l_{ct}	1 if same organ's cell, 0 otherwise
d	pairwise distance between latent embeddings

Model learns biologically meaningful concepts



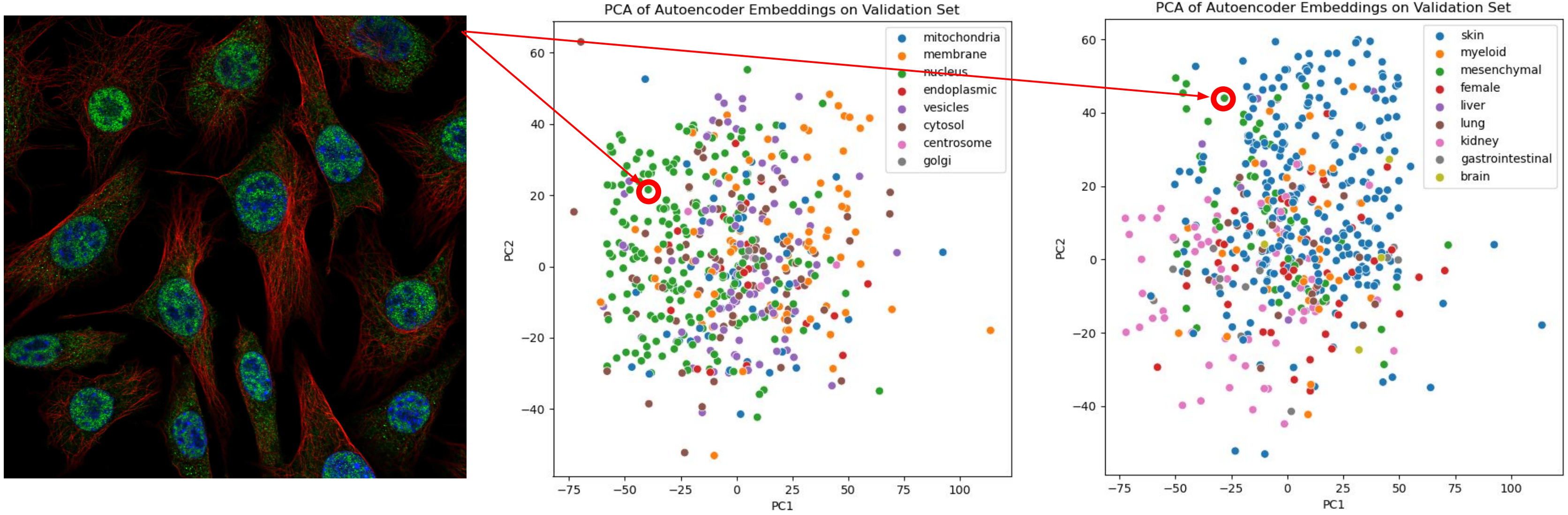
PCA of Autoencoder Embeddings on Validation Set



PCA of Autoencoder Embeddings on Validation Set

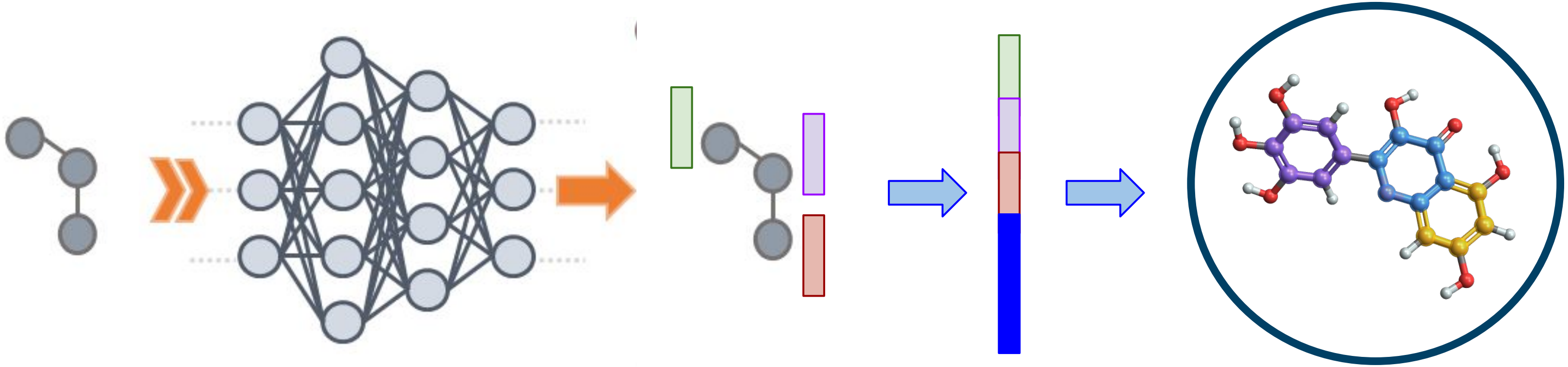
Case study: G-protein coupled receptor kinase 3

- Small molecule drug target for cardiovascular and metabolic disease
- Primary aggregates is nuclei of mesenchymal cells (connective tissue)



Algorithm overview

Conditional Generation



Molecules exhibit some biological specificity

	Control	Experimental	p-value
Hydrophobicity (logP)	2.55 / 0.92	2.21 / 1.27	< 0.0001

- Lower logP values are ideal for absorption, especially in connective tissue

Limitations

Difficulty of measuring biological specificity on generated molecules and cost of VAE training

Conclusions

Images are a promising modality to represent biological context for molecular design

Future Work

Integrating pre-trained protein embeddings could improve interpretability of latent space

References

[1] https://arxiv.org/pdf/2211.02660.pdf

[2] https://www.nature.com/articles/s41467-022-28526-y

[3] https://www.biorxiv.org/content/10.1101/2021.03.29.437595v2.full

[4] https://openaccess.thecvf.com/content/CVPR2022/html/Liu_Multi-Marginal_Contrastive_Learning_for_Multi-Label_Subcellular_Protein_Localization_CVPR_2022_paper.html