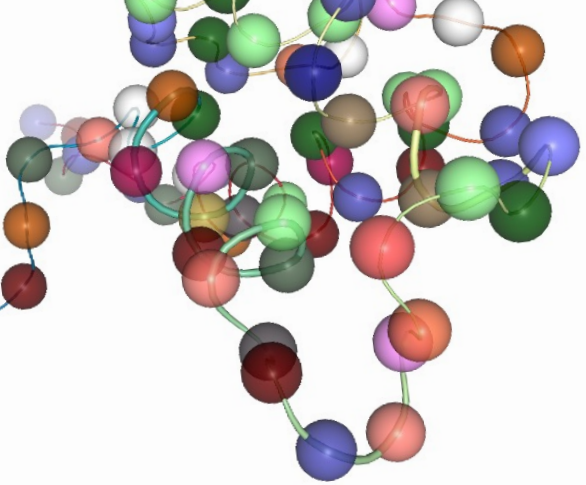
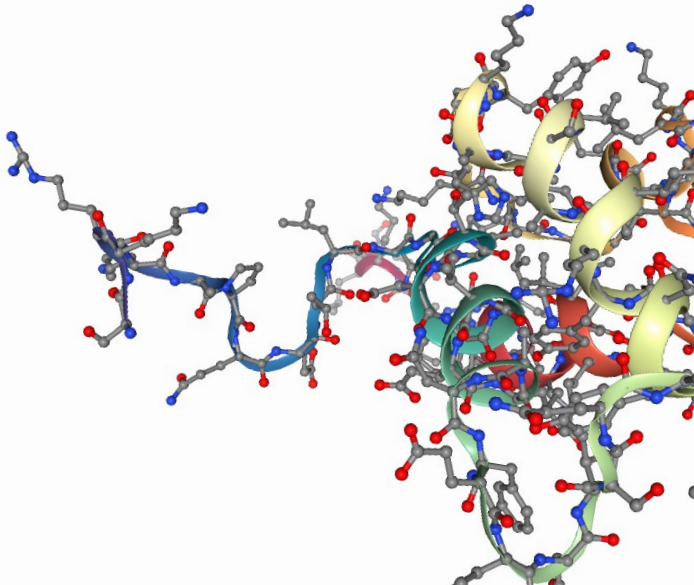


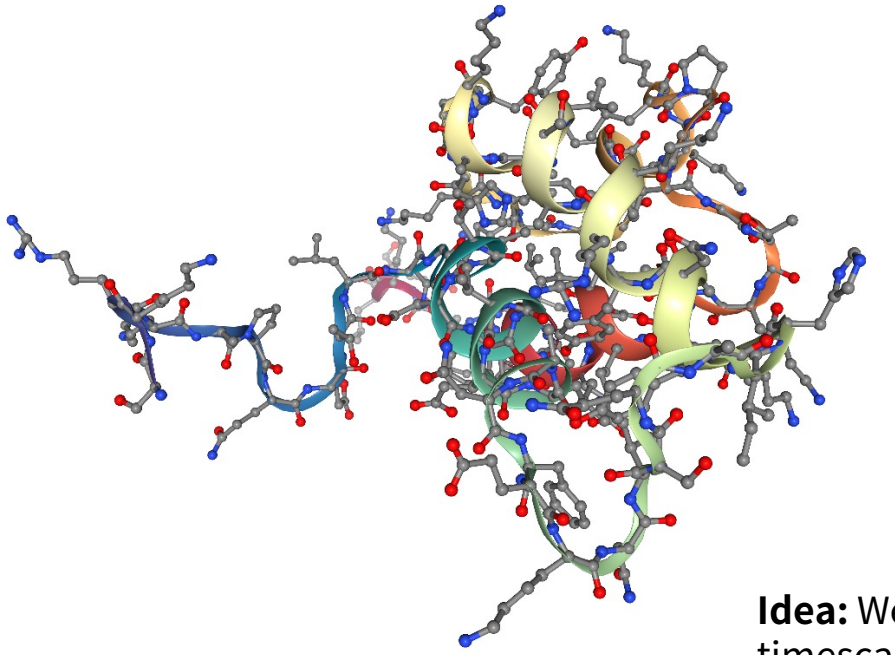
# Chemically Transferable Generative Backmapping of Coarse-Grained Proteins

Soojung Yang, Rafael Gómez-Bombarelli

MIT

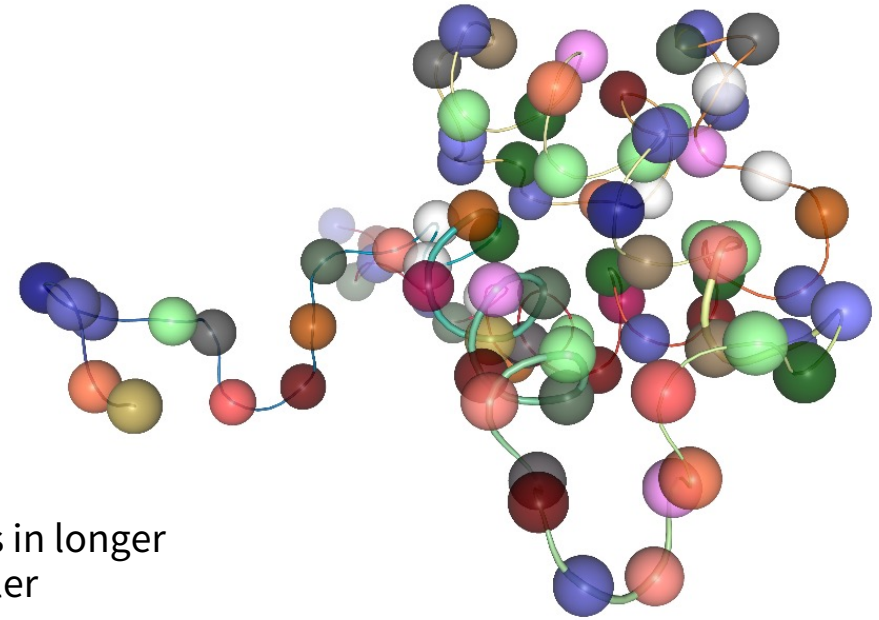


# Coarse-graining and backmapping



Backmapping

Coarse-graining

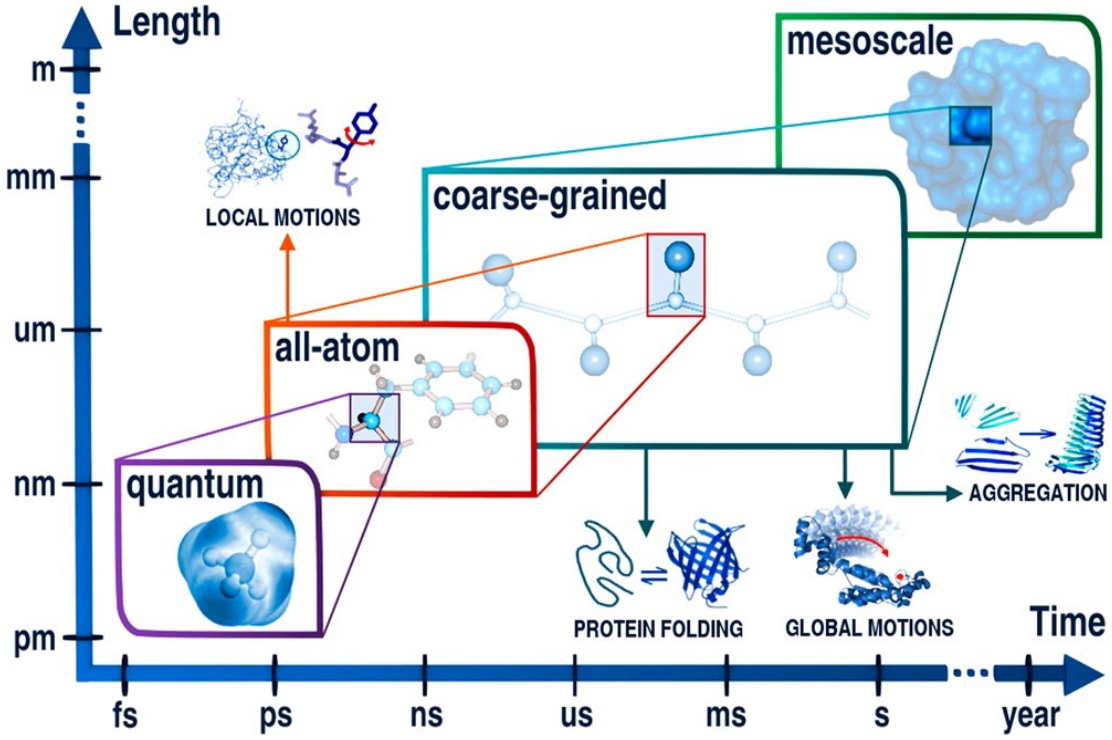


**Idea:** We can simulate large systems in longer timescales by making systems simpler

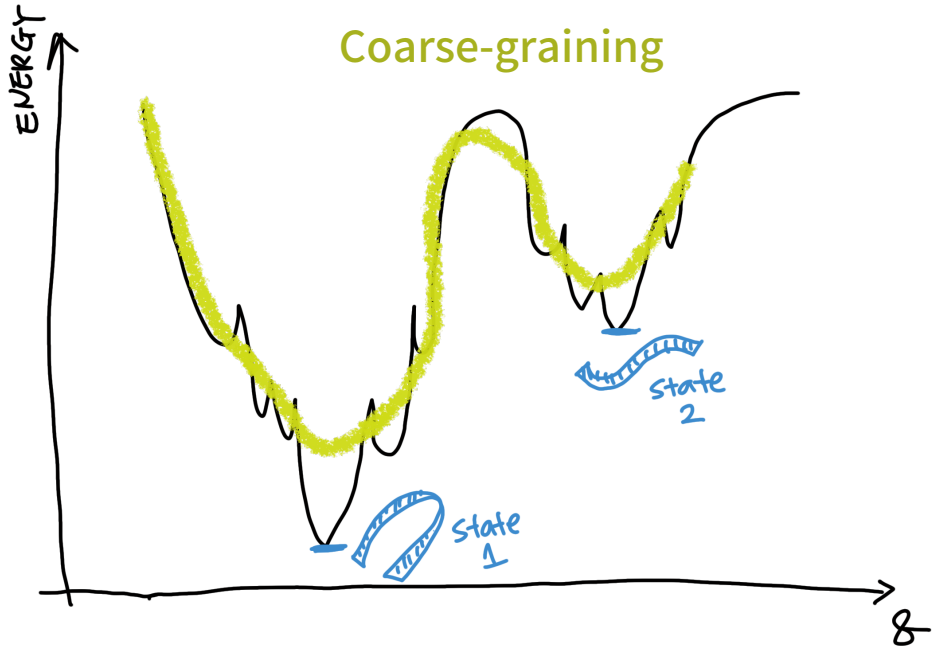
All atom (AA) resolution


Coarse-grained (CG) resolution (residue level CG)

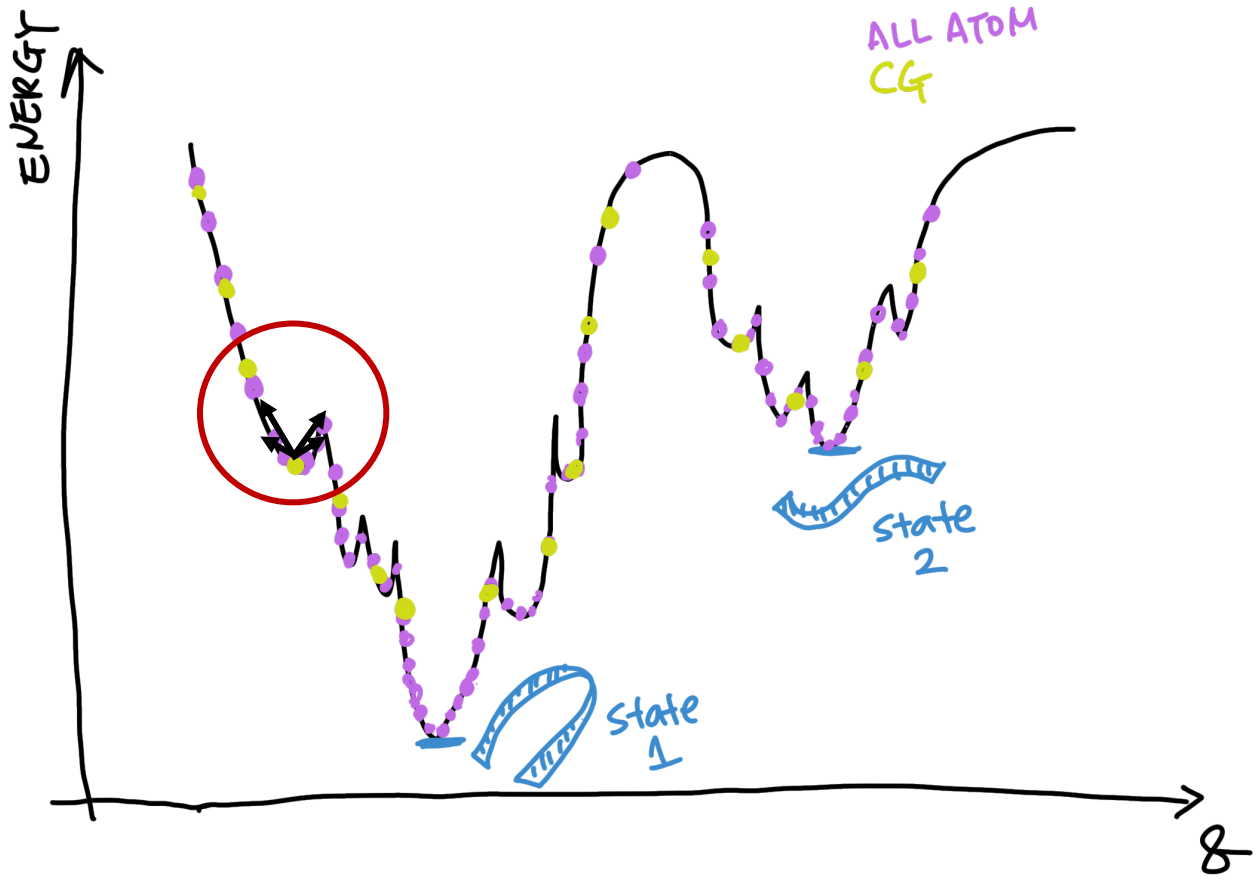
# Coarse-graining allows faster simulations of large biomolecular systems



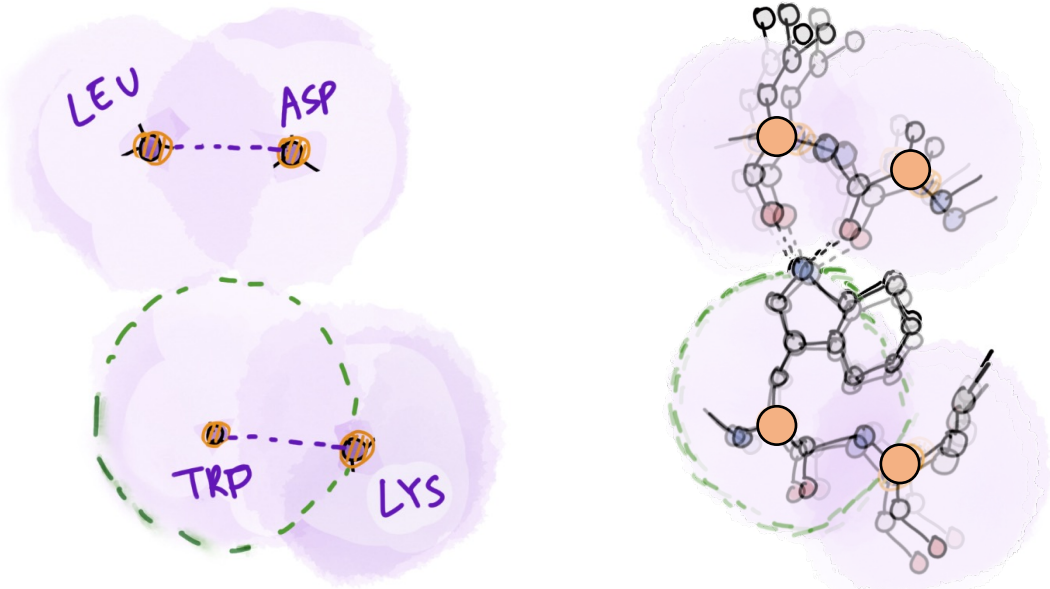

Chem Rev, 2016, doi:10.1021/acs.chemrev.6b00163



# Backmapping is an one-to-many problem that is difficult to solve with deterministic, rule-based methods



# Backmapping as a generative modeling problem




$X$ $x$

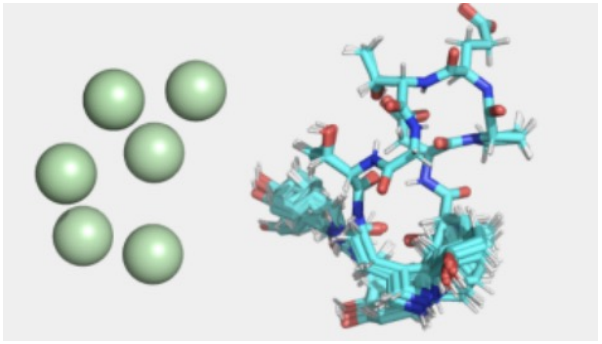
$$P(x|X)$$

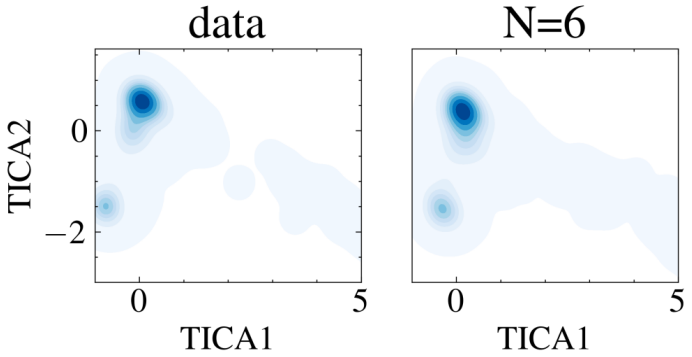
# PoC of generative backmapping with chignolin single chemistry

Wang et al., ICML 2022

**Generative Coarse-Graining of Molecular Conformations**

Wujie Wang [1] Minkai Xu [2 3] Chen Cai [4] Benjamin Kurt Miller [5] Tess Smidt [1] Yusu Wang [4] Jian Tang [2 6 7]
Rafael Gómez-Bombarelli [1]

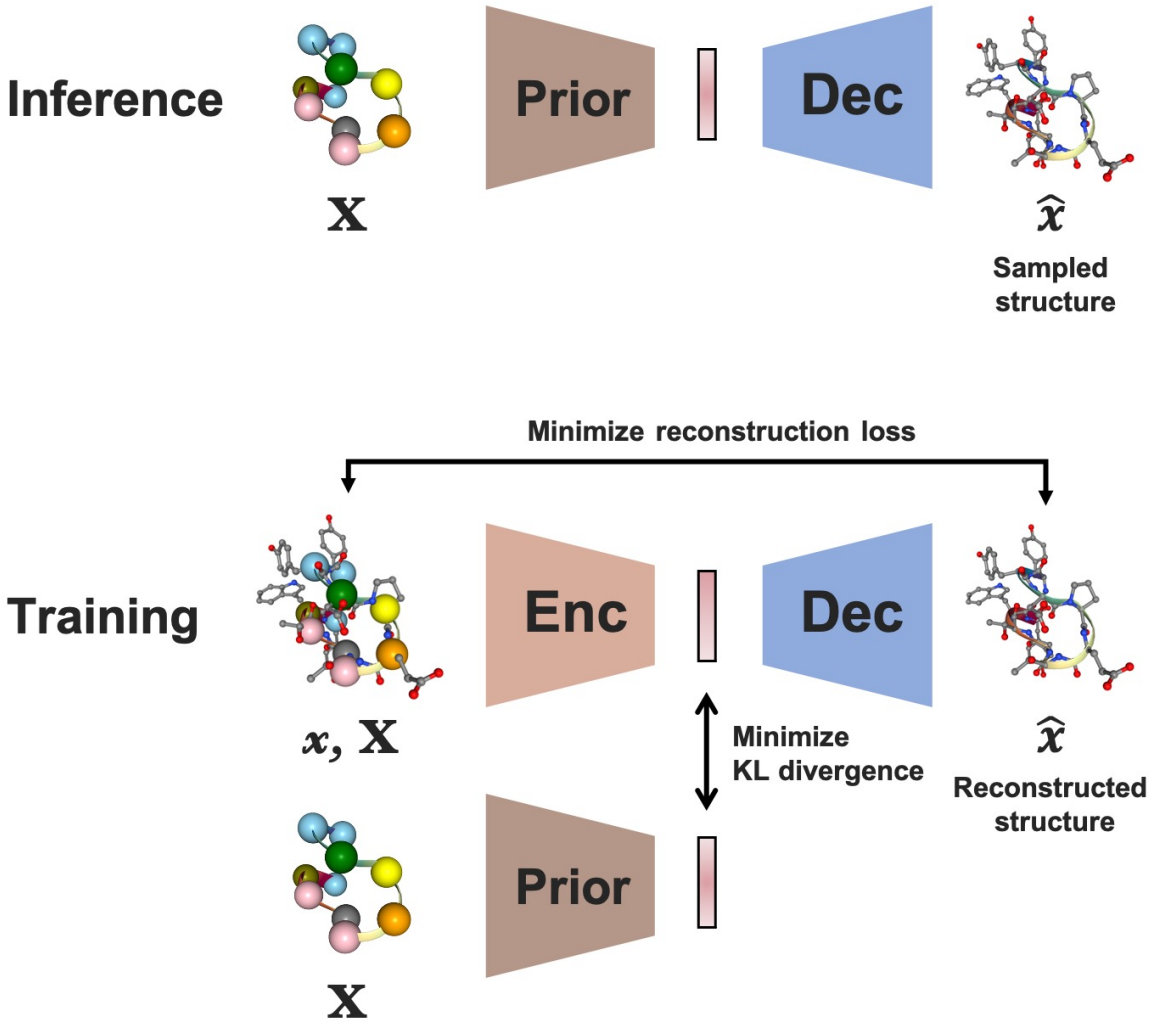




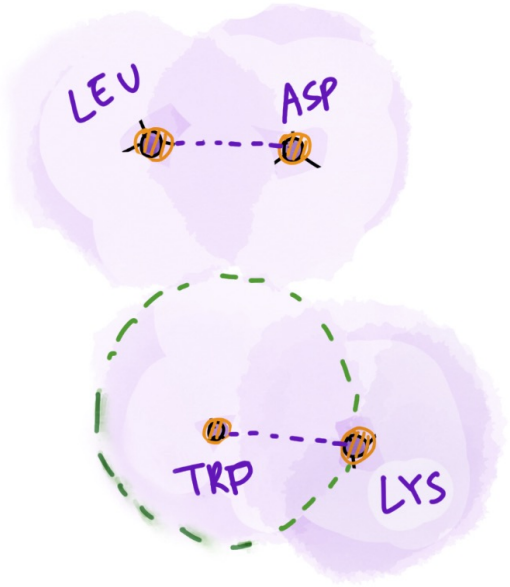
# Conditional VAE framework for backmapping

Wang et al., ICML 2022

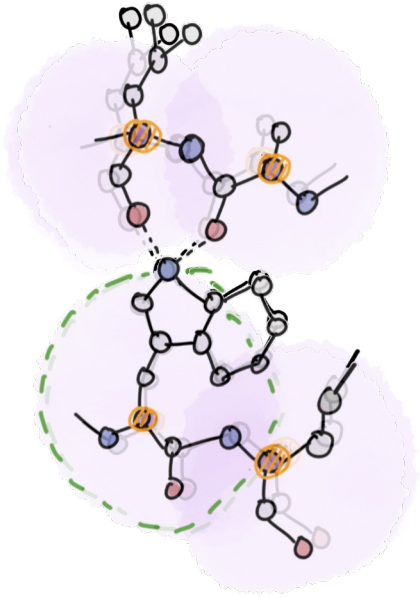
CGVAE



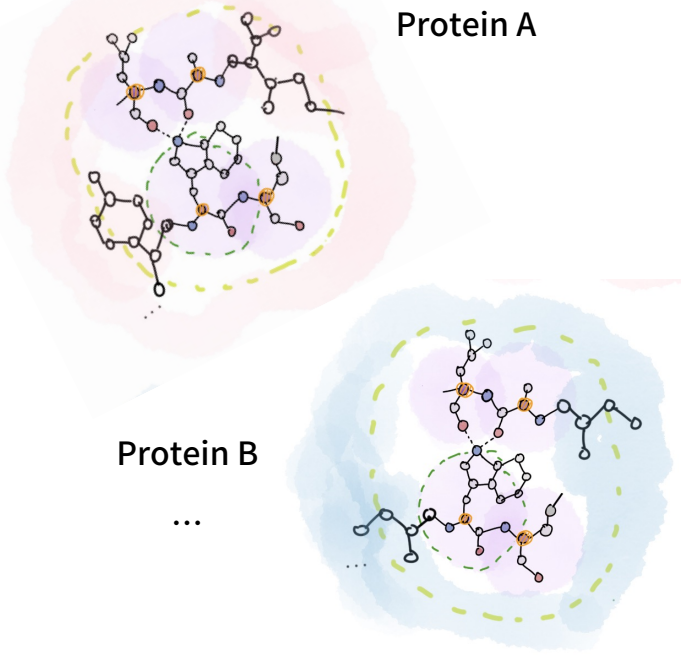
# Transferable backmapping : Inferring from local examples




We don’t know what’s going on underneath the CG level

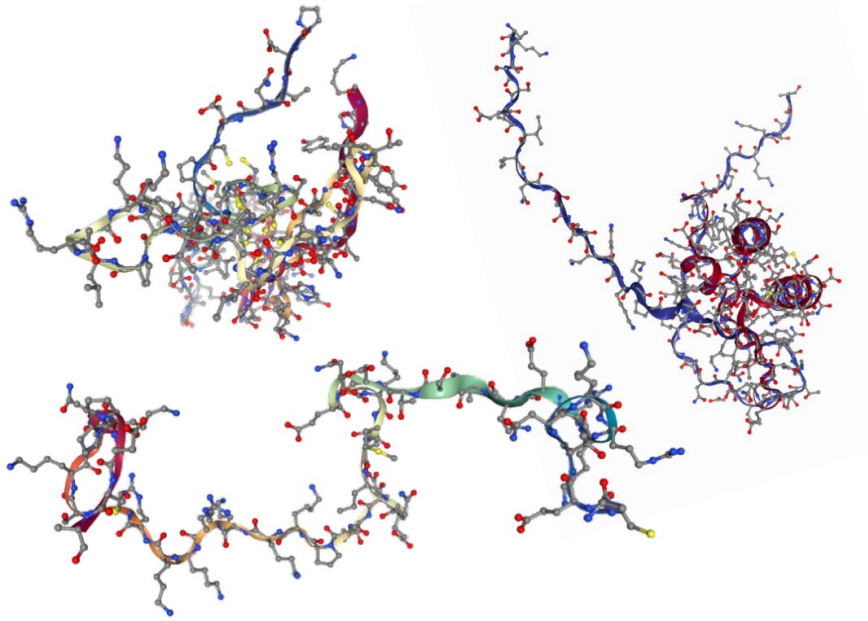


Probably something like this…



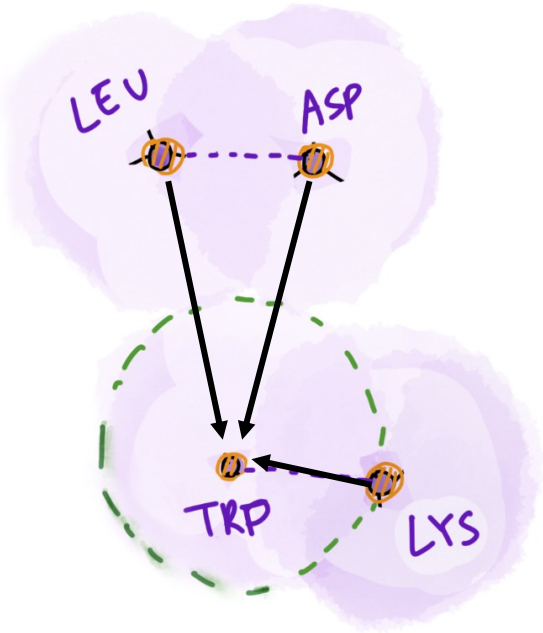

We can infer the local structure from other local examples

# Data to learn transferable backmapping : Protein Ensemble Database (PED)



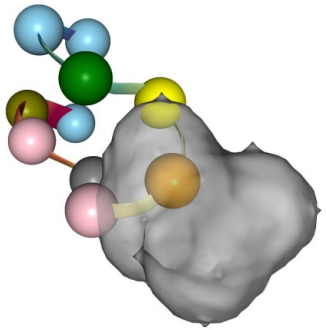
- Database of **protein conformational ensembles**
- **Entry generation**
  - Computational sampling (MD, MC)
  - Ensemble selection

# Transferable backmapping as a local generative modeling problem



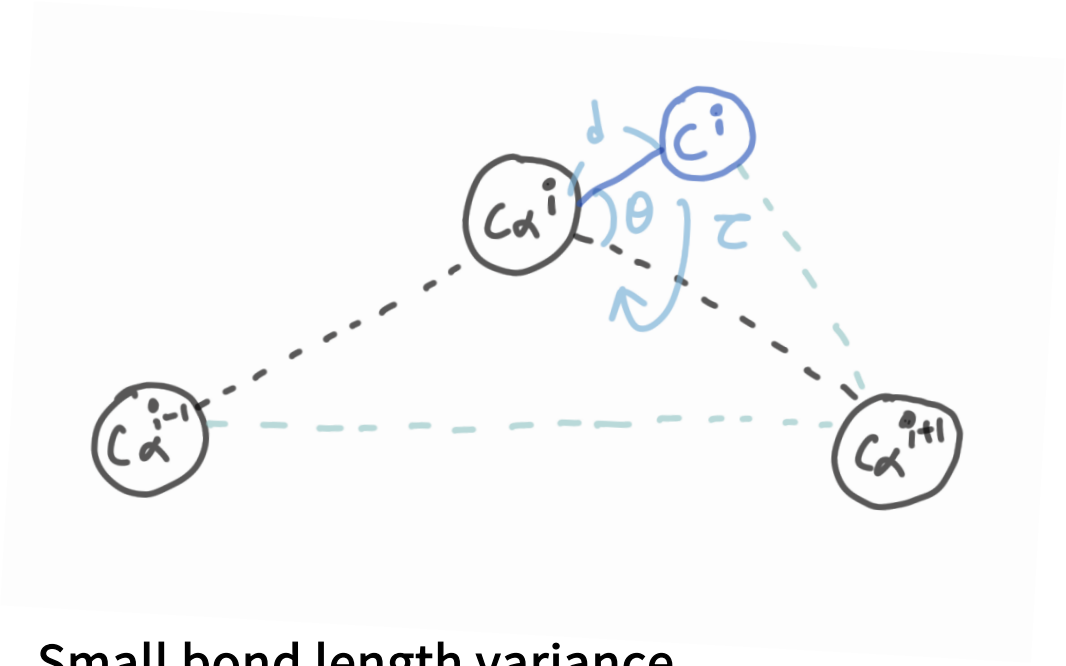

Equivariant message passing
on residues with distance cutoff

Learn orientation
and geometry



Residue-wise
representation & sampling

# Internal coordinate-based backbone reconstruction



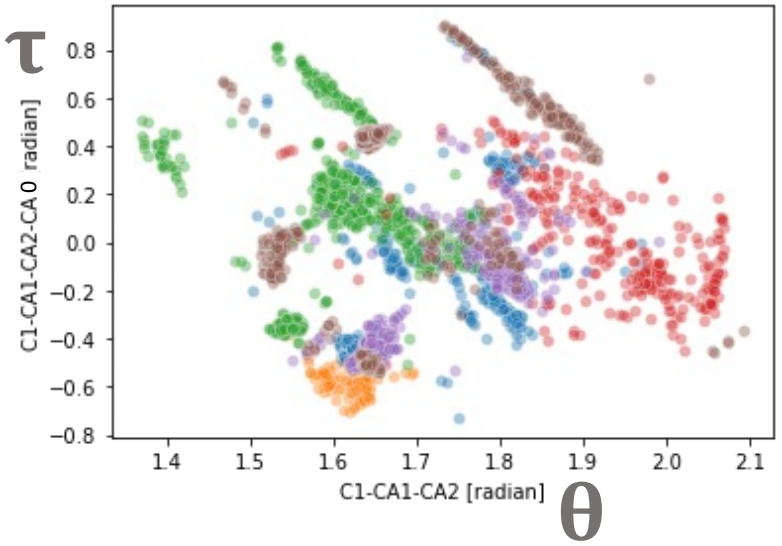

- Output d, $\theta$, $\tau$ of backbone atoms with respect to three adjacent C$\alpha$s

**Small bond length variance**

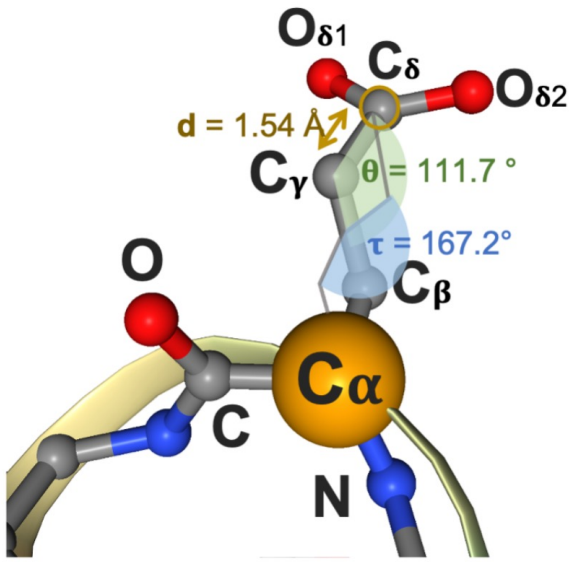
- $\max(d_{C\text{-}CA}) - \min(d_{C\text{-}CA}) = 0.04$ Å
- Predict from the lookup table

**Angle and dihedral are correlated**

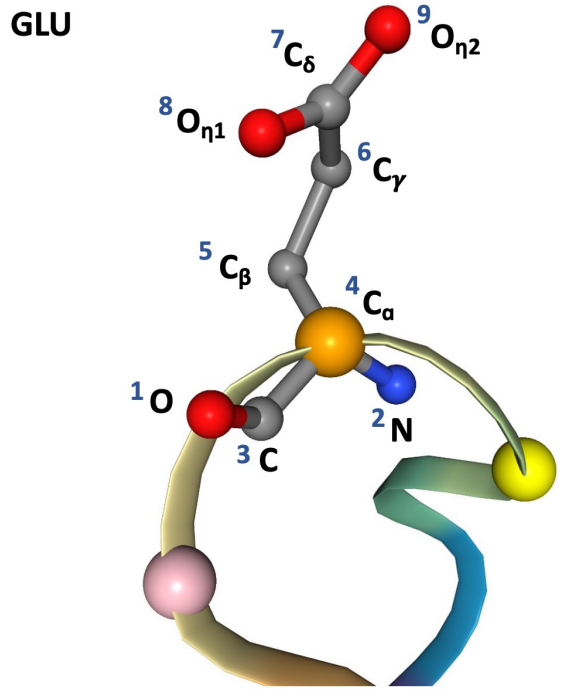
- Jointly predict from the latent variable



# Internal coordinate-based sidechain reconstruction




(b) Sidechain reconstruction from d, $\theta$, $\tau$




- Sequential reconstruction
  - Ex: C**γ** is determined by C**β**, C**α**, and N
- Parallel reconstruction of all residues
  - Ex: Reconstruct all C**γ**s at once

# Learning objectives

Supervision over internal coordinates

- Bond length
- Bond angle
- Torsion angle

Supervision over Cartesian coordinates

- RMSD
- Steric clash

$$\underbrace{\frac{1}{|B|}\sum_{b\in B}(b-\hat{b})^2}_{L_{\text{bond}}}+\underbrace{\frac{1}{|A|}\sum_{\theta\in A}\sqrt{2(1-\cos(\theta-\hat{\theta}))+\epsilon}}_{L_{\text{angle}}},$$

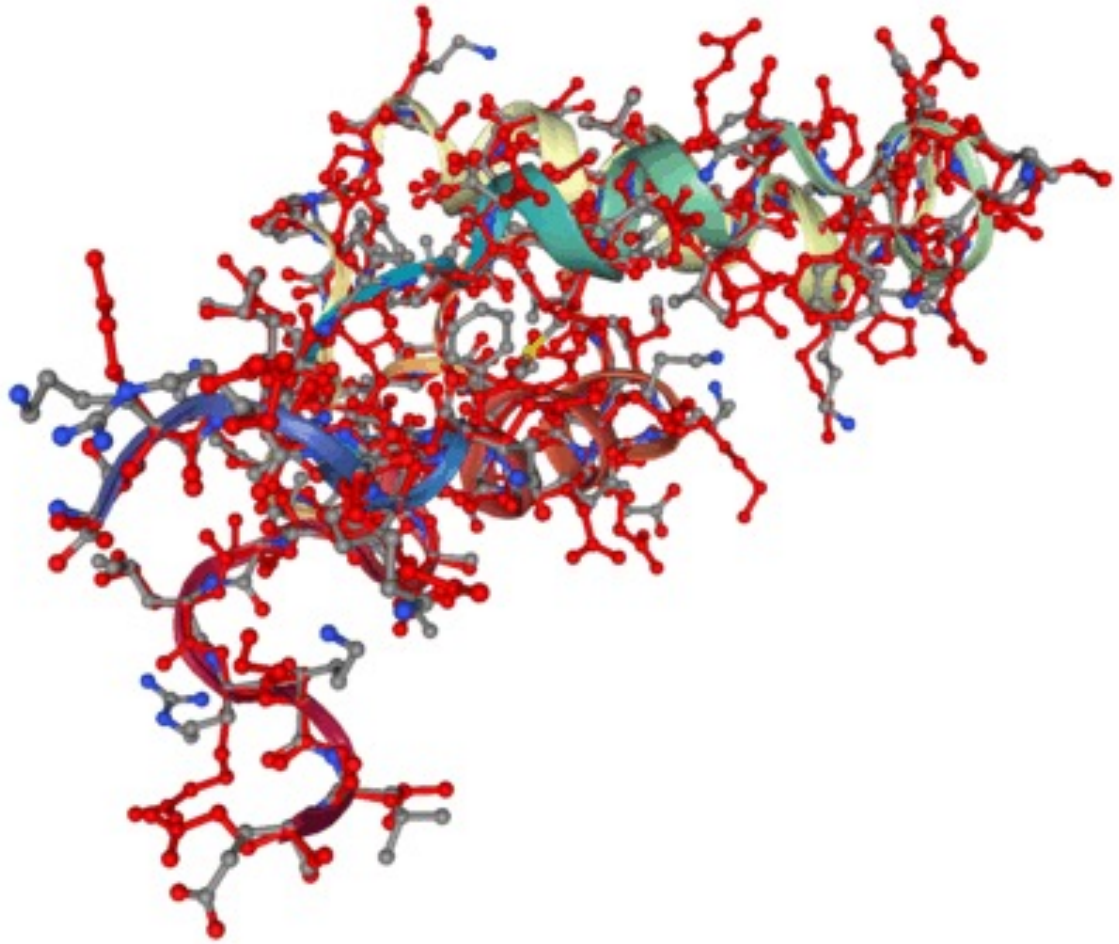
$$L_{\text{torsion}} \coloneqq \frac{1}{|T|}\sum_{\tau\in T}\sqrt{2\times(1-\cos(\tau-\hat{\tau}))+\epsilon}$$

$$L_{\text{steric}} \coloneqq \sum_{\mathbf{x}\in\mathcal{N}}\sum_{\mathbf{y}\in\mathcal{B}_r(\mathbf{x})}\max(2.0-||\mathbf{x}-\mathbf{y}||_2^2, 0.0)$$

# Model performance

**PED00055**
**N=10 sampling**

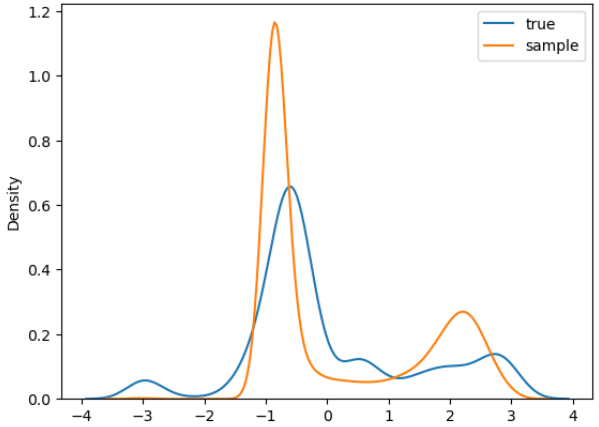
**Red:** Reference AA structure
**Color coded:** sampled structures



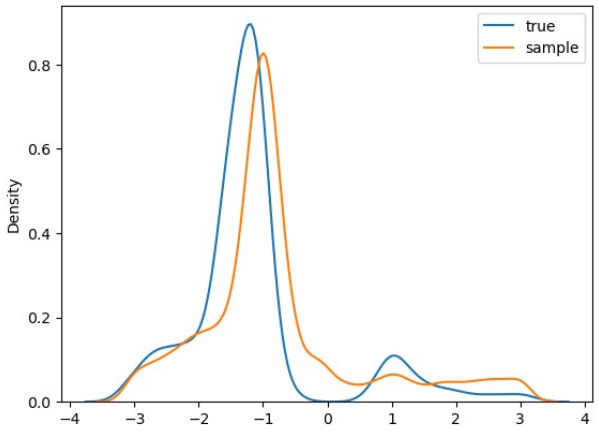
# Ground truth and sampled distributions of backbone and side chain dihedral angles

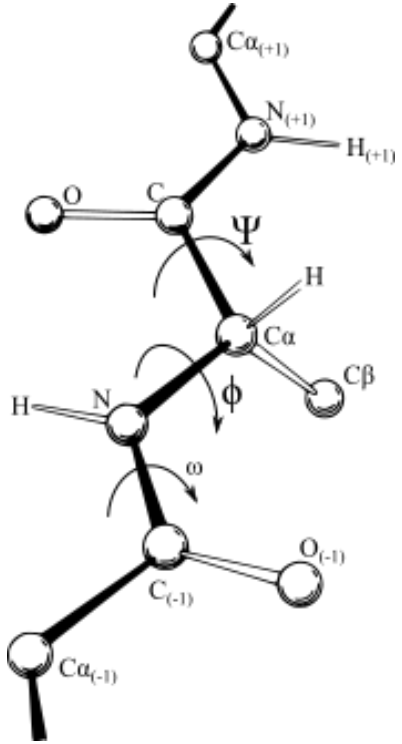
**PED00055**
**N=10 sampling**

**Backbone Psi**



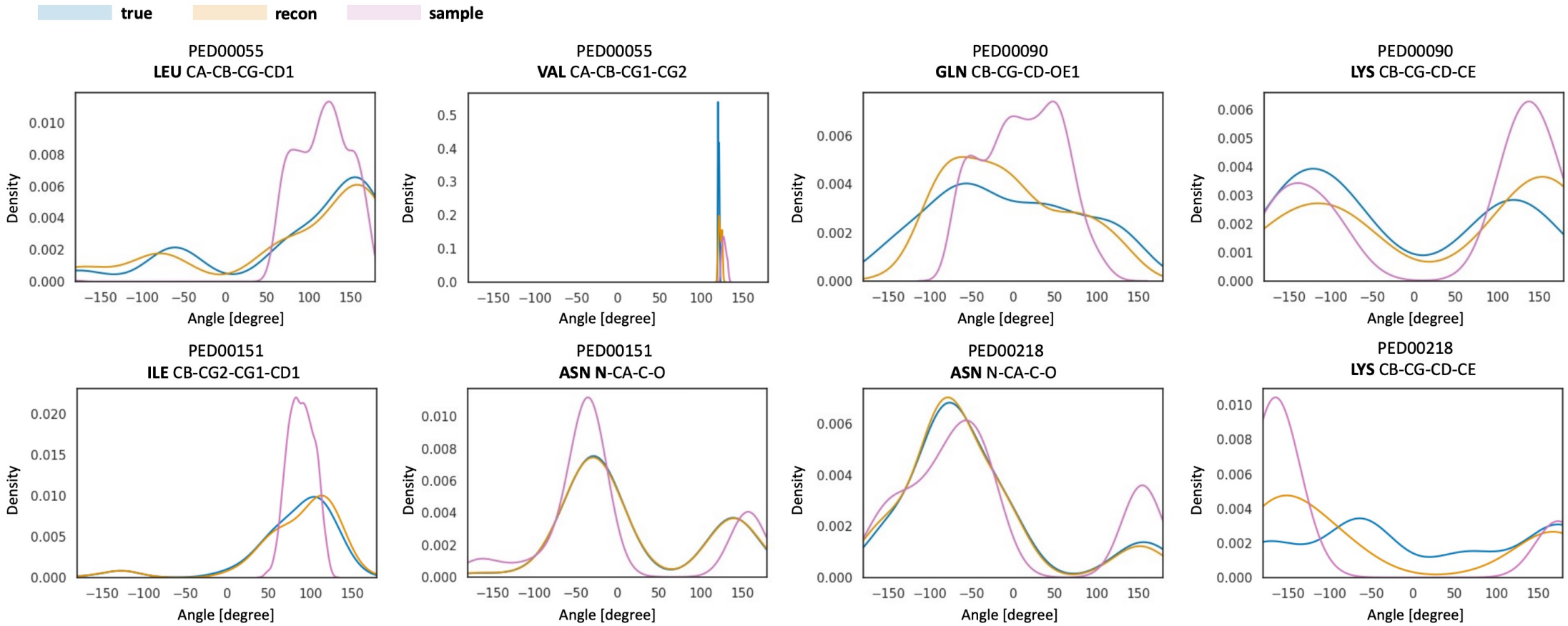

**Backbone Phi**




Backbone min RMSD: approx. 0.5 Å

# Ground truth and sampled distributions of backbone and side chain dihedral angles




Sidechain min RMSD averaged across all amino acid types : approx. 1.3 Å

# Thank you!