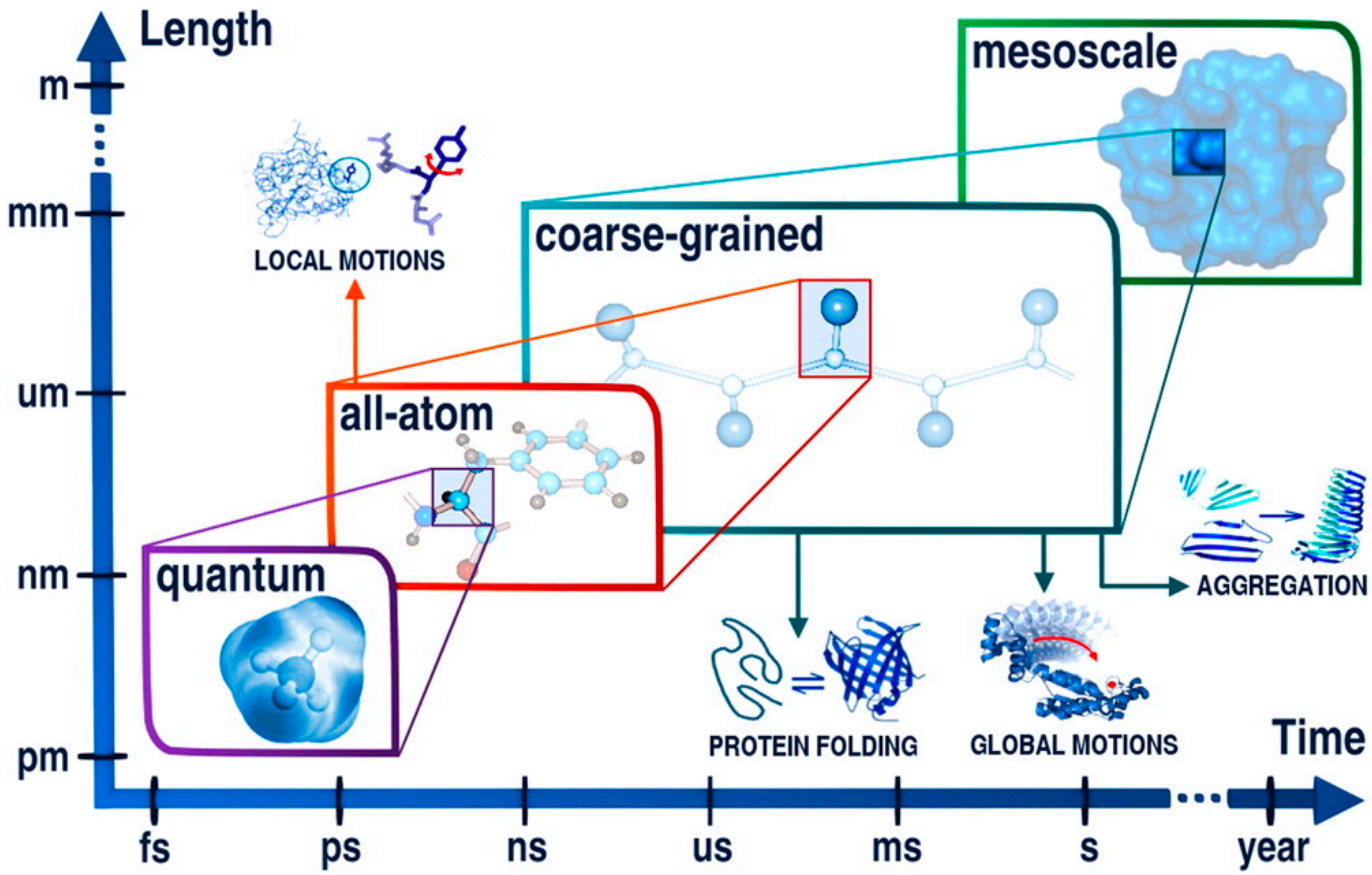
# Generative Coarse-Graining of Molecular Conformations

Wujie Wang, Minkai Xu, Chen Cai, Benjamin Kurt Miller, Tess Smidt, Yusu Wang, Jian Tang, and Rafael Gomez-Bombarelli

ICML
International Conference On Machine Learning

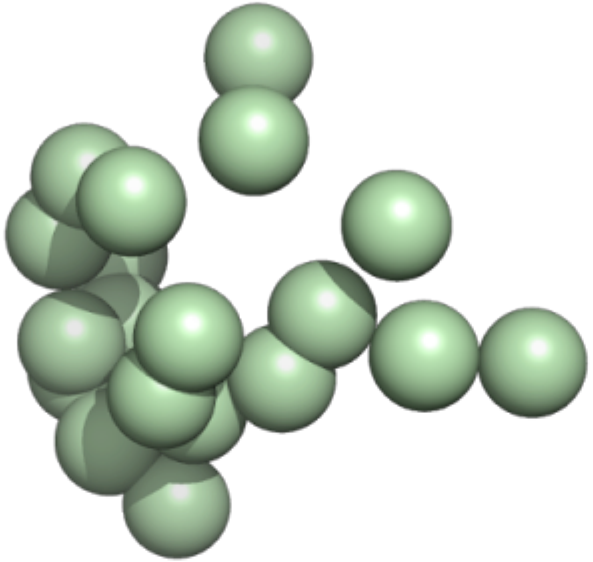
# Coarse-Graining for Molecular Modeling

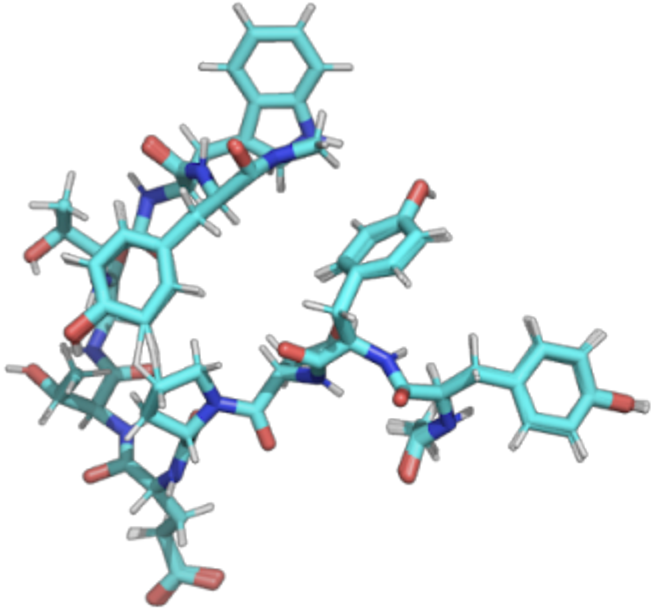



- Coarse-grained modeling can greatly accelerate speed of molecular dynamics
- However, recovering fine-grained coordinates is challenging, due to loss of information.

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., & Kolinski, A. (2016). Coarse-grained protein models and their applications. Chemical reviews, 116(14), 7898-7936.

# Super-resolution for molecular geometries





$X \in \mathbb{R}^{N \times 3}$ $\qquad$ $x \in \mathbb{R}^{n \times 3}$

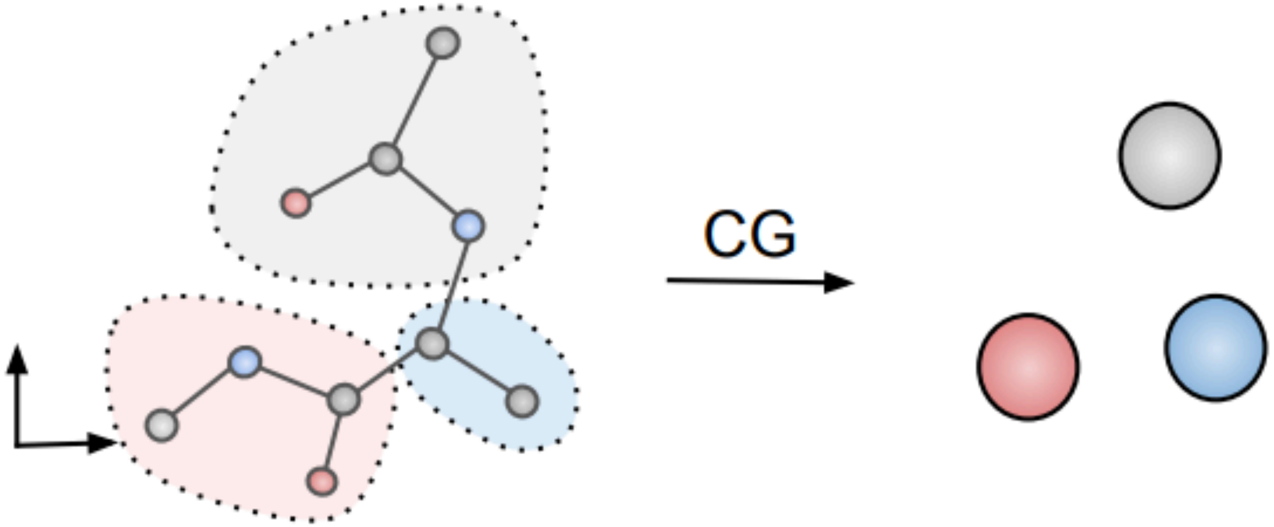
# Our contributions

- We proposed a generative modeling framework (**CGVAE**) for the backmapping task using geometric deep learning, i.e. modeling **p(x|X)**

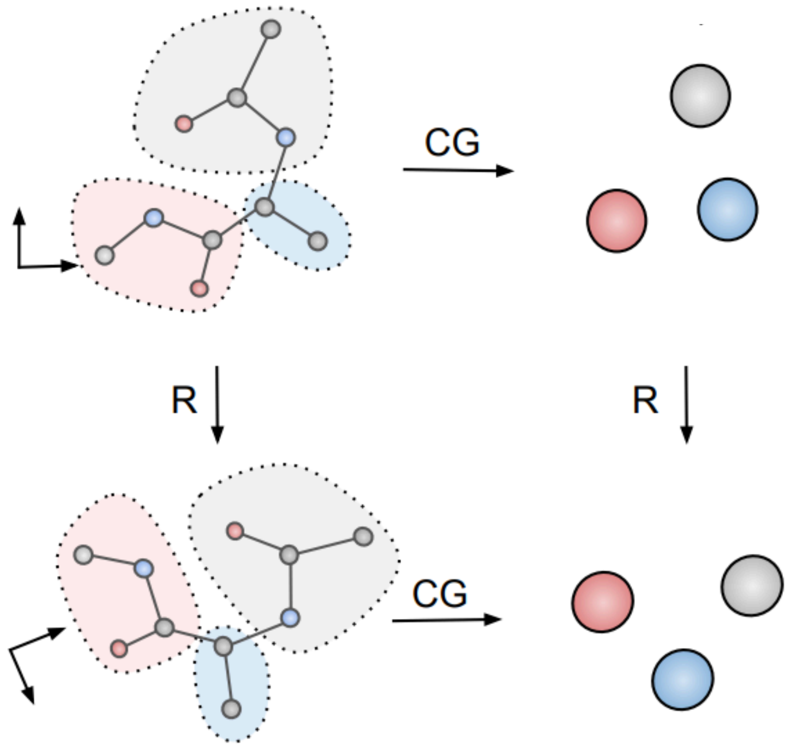
  - **Generality** → A model with geometric data representations that work for arbitrary mapping and resolution (it is designed to work very coarse representations)
  - **Geometric constraints** → We derive the geometric constraints for backmapping, we explicitly incorporate these constraints in our model
  - **One-to-many stochastic map** → Explicitly model p(x|X)

  - Evaluation metrics and protocols

# Particle-based Coarse-Graining

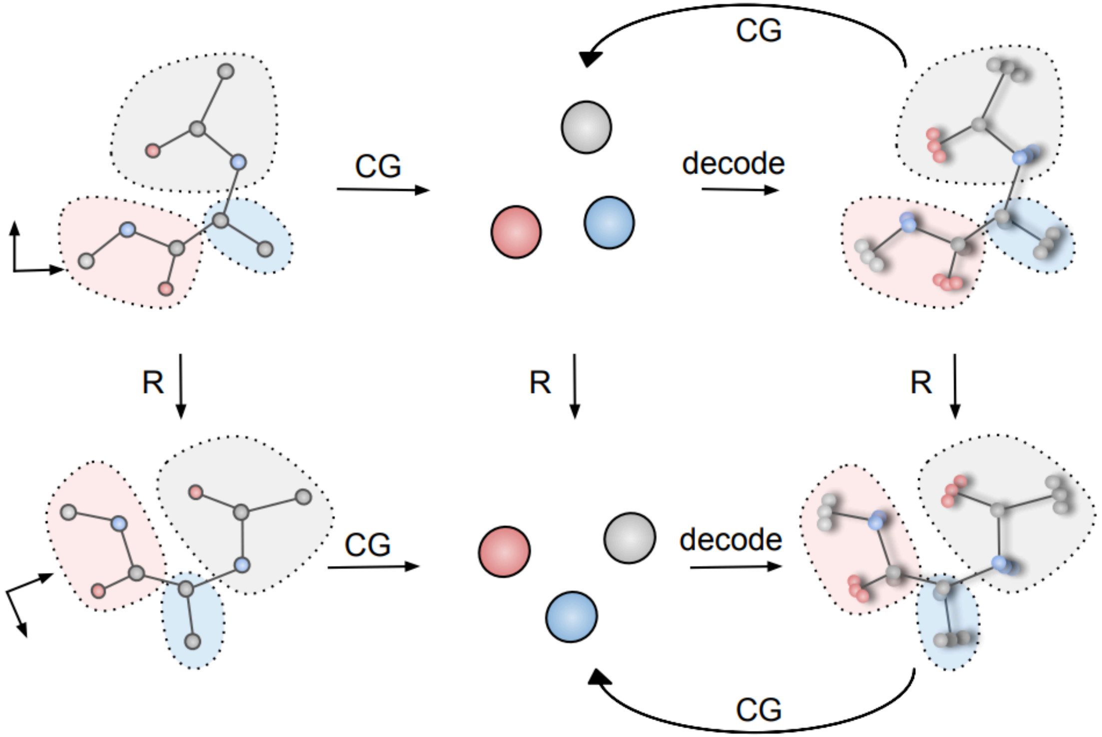


**The system rotates, reflects and translates the same way before and after coarse-graining.**

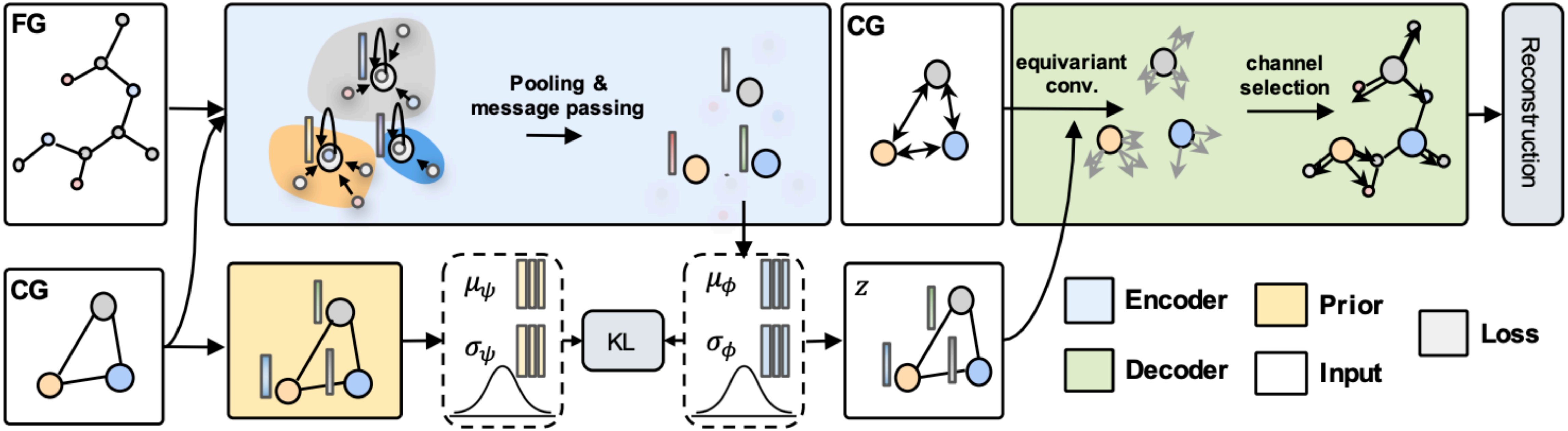


**The backmapping transformation needs to be E(3) equivariant, and produce Fine-grained geometries that are compatible with the Coarse-grained geometries.**

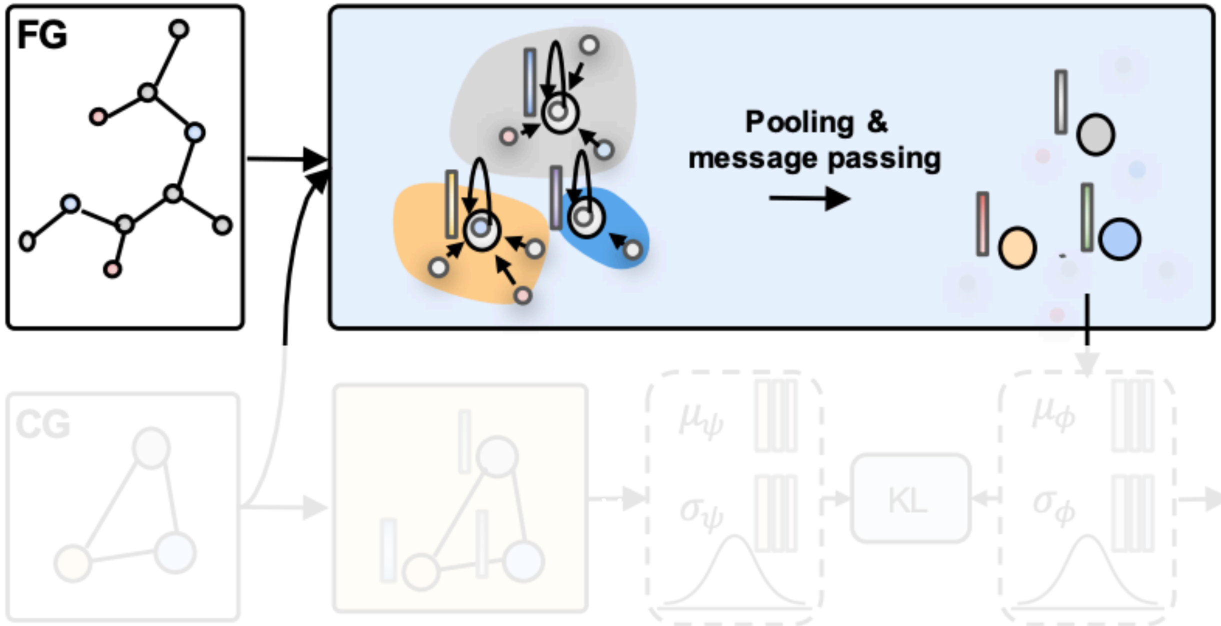
$$\textbf{R1}.\ M\widetilde{x} = M\mathrm{Dec}(X) = X.$$
$$\textbf{R2}.\ \mathrm{Dec}(QX + g) = Q\mathrm{Dec}(X) + g.$$



# Model Design



# Encoder **q(z|x, X)**




**MPNN update**

Radial Basis Expansion

inter-atomic distance

FG features $\in \mathbb{R}^F$

$$h_i^{t+1} = \mathrm{Update}\Big(\sum_{j\in N(i)} \mathrm{Msg}(h_i^t, h_j^t, \mathrm{RBF}(d_{ij})),\ h_i^t\Big)$$

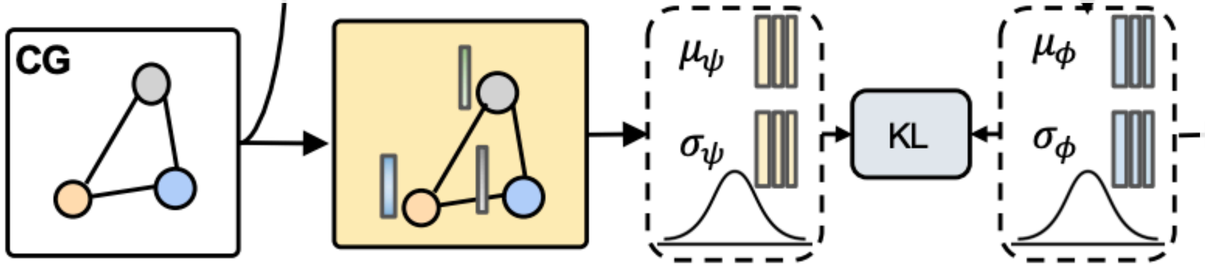
$$\widetilde{H}_I^t = \mathrm{Update}\Big(\sum_{i\in C_I} \mathrm{Msg}\left(h_i^{t+1}, H_I^t, \mathrm{RBF}(d_{iI})\right),\ H_I^t\Big)$$

CG features $\in \mathbb{R}^F$

$$H_I^{t+1} = \mathrm{Update}\Big(\sum_{J\in N(I)} \mathrm{Msg}(\widetilde{H}_I^t, \widetilde{H}_J^t, \mathrm{RBF}(d_{IJ})),\ H_I^t\Big)$$

$$q_\phi(z|x,X) = \prod_I^N \mathcal{N}(\mu_{\mathbf{MPNN},I}(x,X), \sigma_{\mathbf{MPNN,I}}(x,X))$$

# Prior p(z|X)

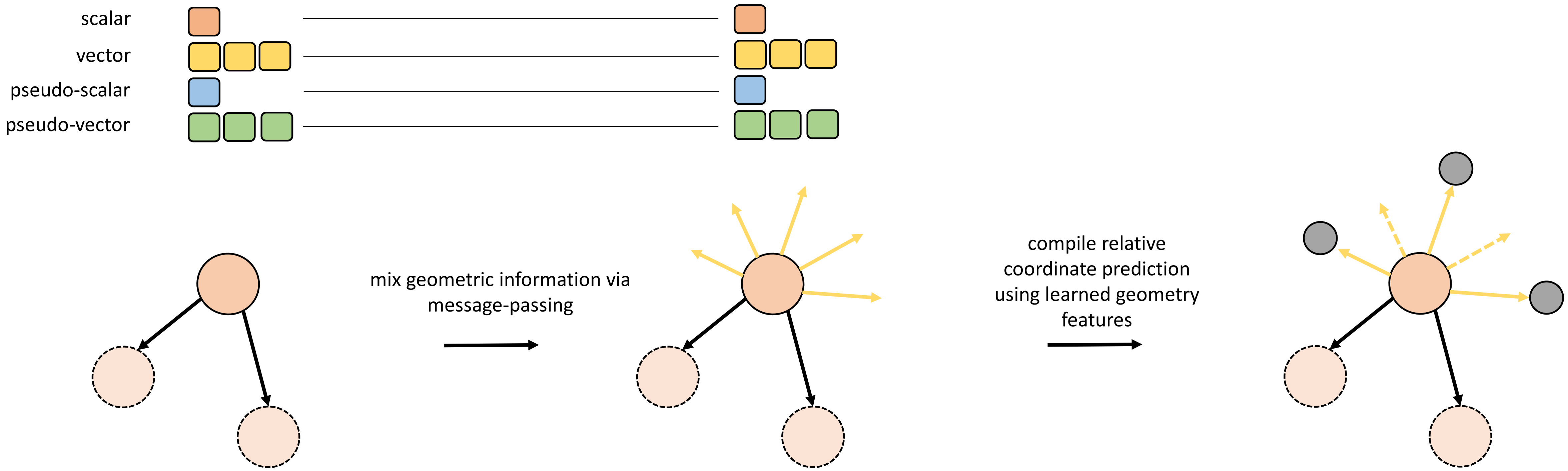



**MPNN update**

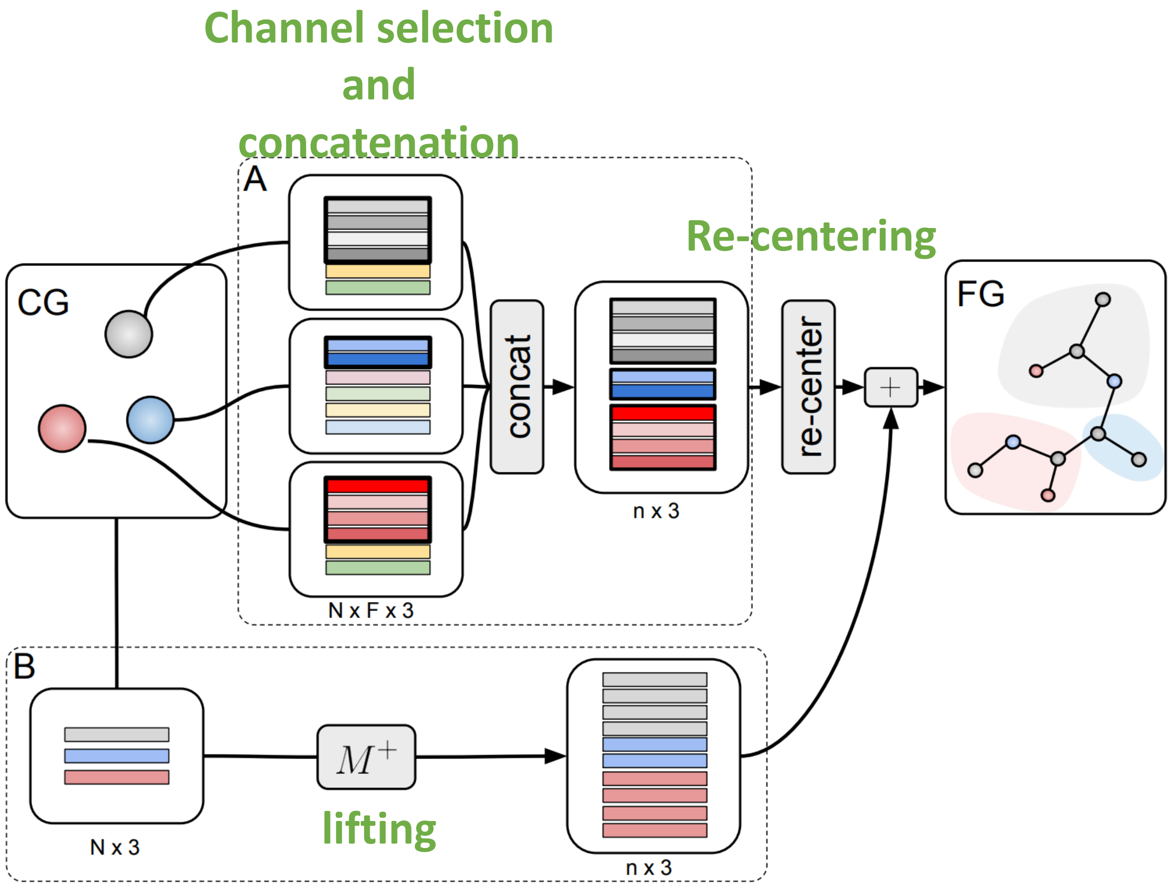
$$H_I^{t+1} = \mathrm{Update}\Big(\sum_{J\in N(I)} \mathrm{Msg}(\widetilde{H}_I^t, \widetilde{H}_J^t, \mathrm{RBF}(d_{IJ})),\ H_I^t\Big)$$

$$p_\psi(z|X) = \prod_I^N \mathcal{N}(\mu_{\mathbf{MPNN},I}(X), \sigma_{\mathbf{MPNN,I}}(X))$$

# Decoder **p(x|X, z)**

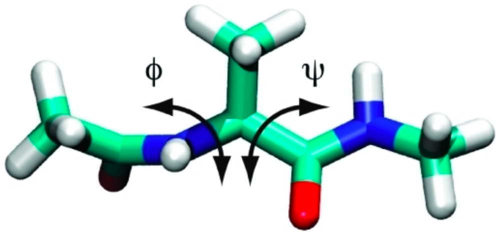


# Compile prediction for geometries

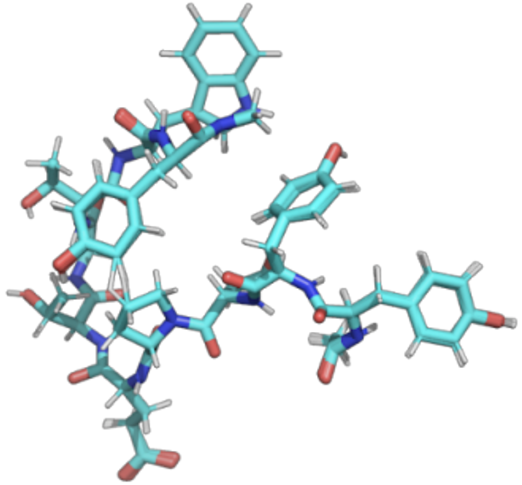


# Experiment – datasets, metrics

- **Datasets**
  - **Alanine dipeptide trajectory:** 22 atoms**,** a classical molecular system for benchmarking enhanced Sampling
  - **Chignolin trajectory:** 175 atoms, a mini-protein that features folded and unfolded states
- **Metrics**
  - **Reconstruction RMSD** – evaluates the model's capacity to reconstruct fine-grained geometries. **The lower the better**
  - **Sampling RMSD** – compares the average RMSD between generated structures and reference structures. **The higher the better.**
  - **Graph Validity** – measures how well the generated FG geometries preserve the original chemical bond graphs. **The lower the better**
- **Coarse-grained mapping generations:**
  - We use Coarse-grained auto-encoders[1] to generate a mapping for our experiments. Other mapping also works, even random mappings.

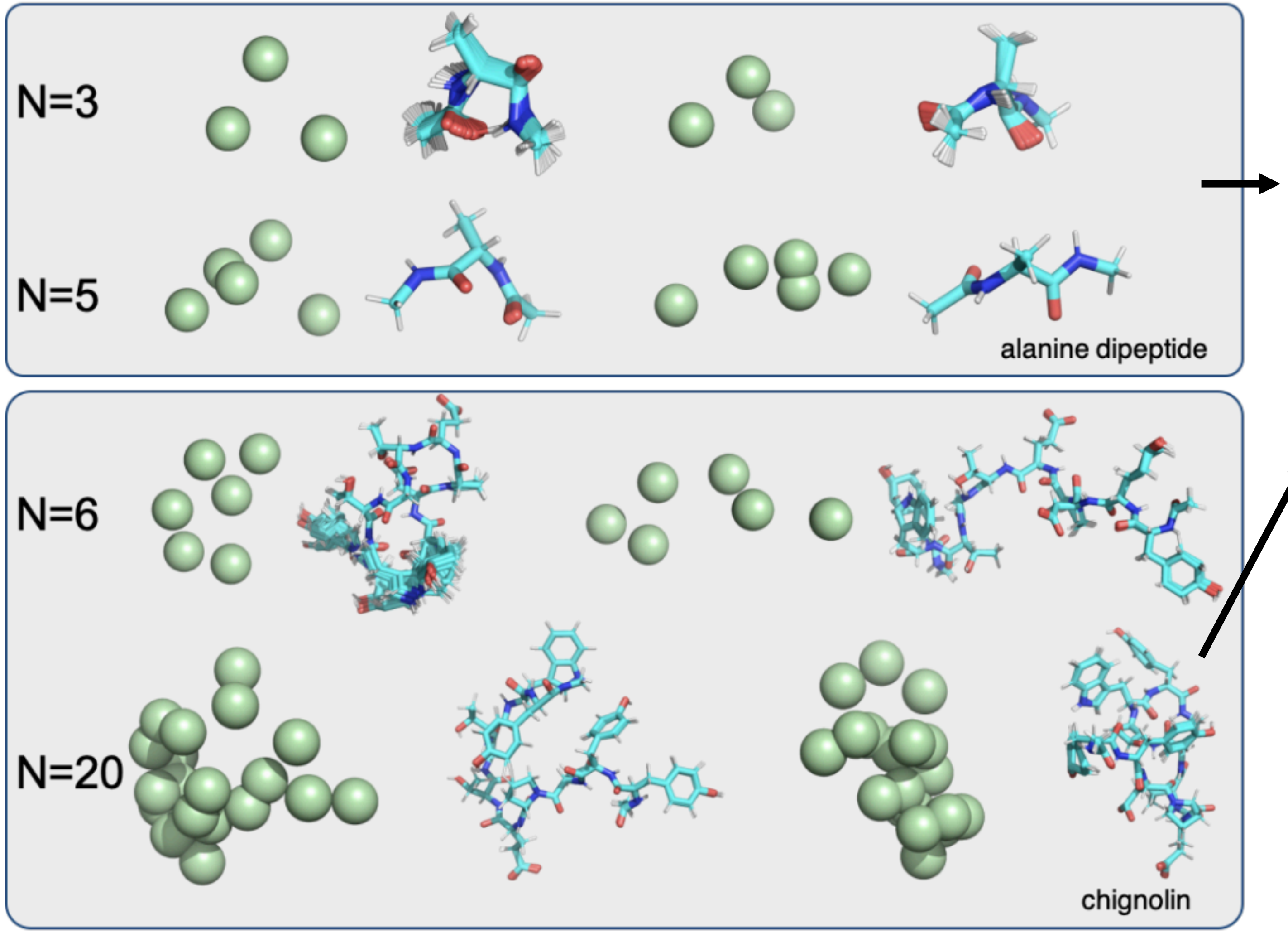


Alanine dipeptide


Chignolin

Wang, W., & Gómez-Bombarelli, R. (2019). Coarse-graining auto-encoders for molecular dynamics. *npj Computational Materials*, *5*(1), 1-9.

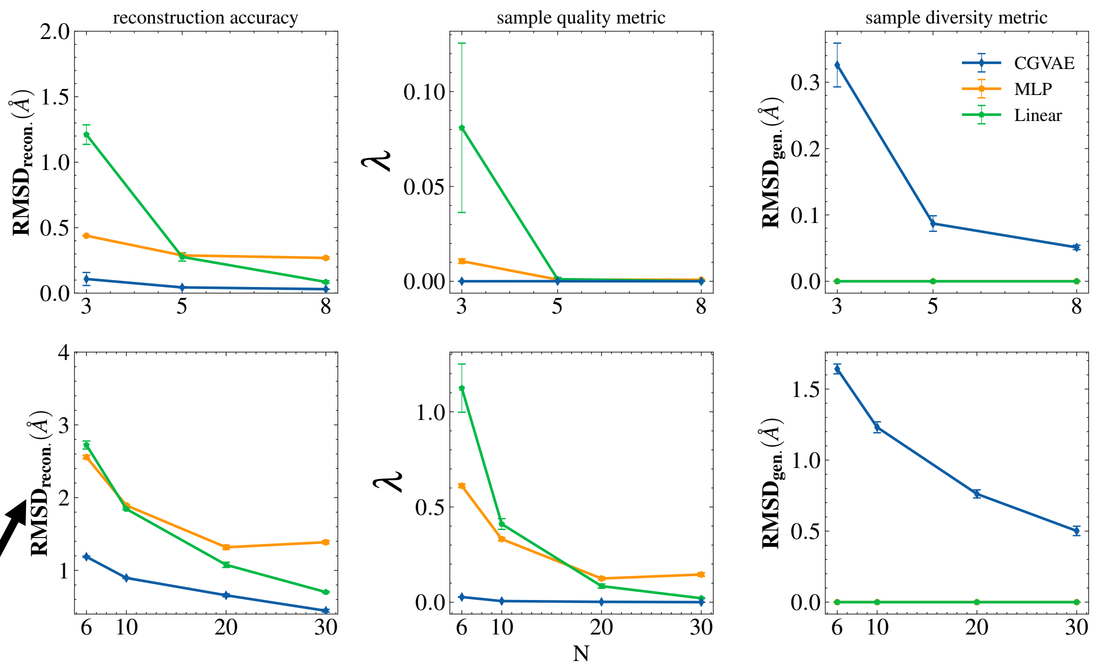
# Experiments




Reconstruction Accuracy ↓

Graph Validity ↓

Sample diversity↑




- We benchmark our model at different level of coarsening
- Accurate backmapping is achieved with even very coarse representations – much coarser than usual CG mapping choices