Probing Chirality and its Latent Space Organization in E(3) Equivariant Boltzmann Generators for Alanine Dipeptide

Seh Na Mellick[1,2], Sanjyot Shenoy[1,3]
[1]Joint Carnegie Mellon-University of Pittsburgh Ph.D. Program in Computational Biology
[2]Ray and Stephanie Lane Computational Biology Department, Carnegie Mellon University
[3]Center of Systems Immunology, University of Pittsburgh

INTRODUCTION

$g_0 \sim p_0(g) \longrightarrow \mathrm{d}g = u_\theta(g,t)\mathrm{d}t \longrightarrow g_1 \sim p_1(g)$

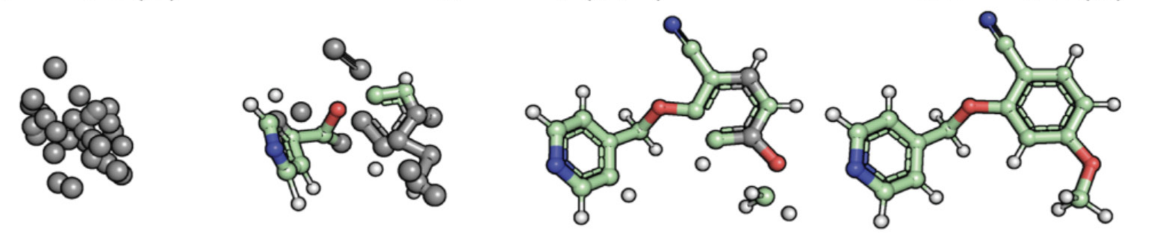
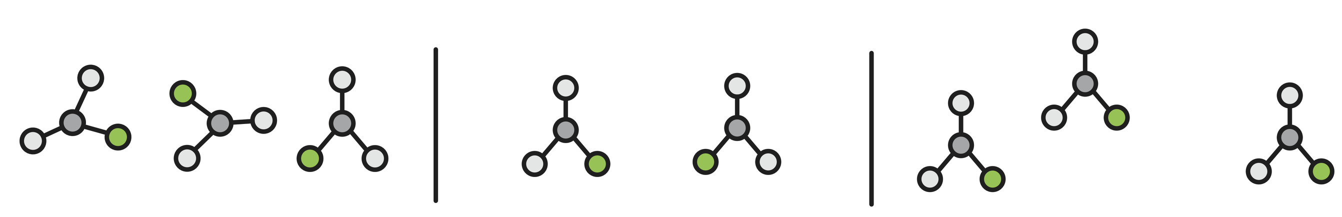


Figure Credit - Exploring Discrete Flow Matching for 3D De Novo Molecule Generation, I. Dunn, D. Koes, NeurIPS 2024

Flow Matching (FM)-based generative modeling has become popular, owing to its efficient design



E(3) equivariant models preserve rotations (left), reflections (middle), and translations (right)

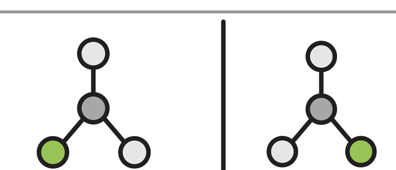
- **Boltzmann Generators** use E(3)-equivariant normalizing flows to directly and efficiently sample molecular equilibrium states governed by the Boltzmann distribution.
- Incorporating physical symmetries via equivariant architectures is crucial for ensuring these generators produce physically realistic molecular structures and learn efficiently.

$x_0 \sim N(0, I)$

E(3)-equivariant CNF

$x_1 \sim \mu(x_1)$

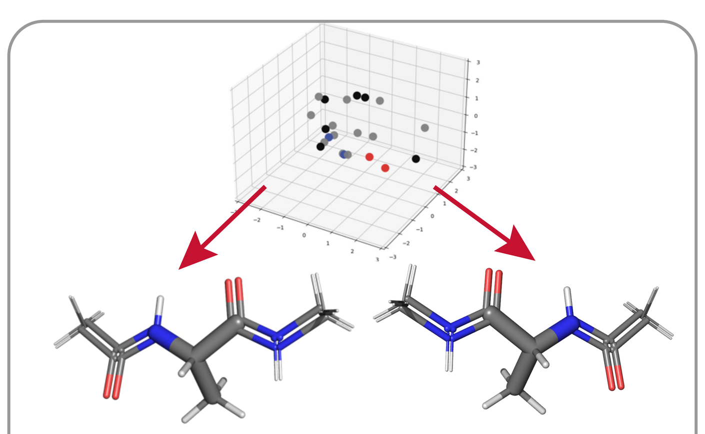
$\mu(x) \propto \exp\left(-\frac{U(x)}{k_B T}\right)$



Chiral molecules are mirror images- identical in mass, distinct in handedness

E(3)-equivariant model architectures do not distinguish between enantiomers because of reflection equivariance

OBJECTIVES

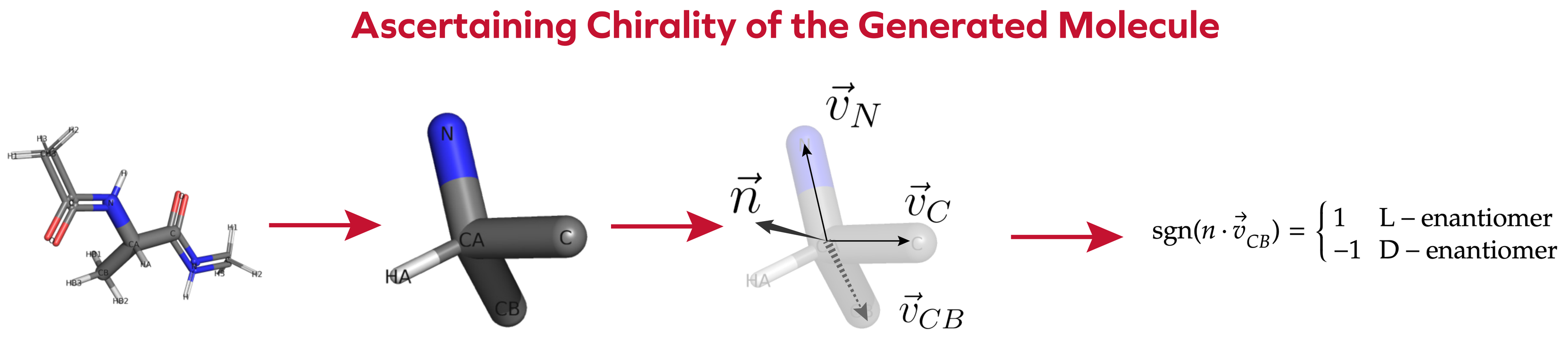


E(3)-equivariant architectures preserve reflections, which in a biological molecule has critical significance. We investigate chirality-related artifacts of the E(3)-equivariant BG-generated samples for Alanine Dipeptide (AD2) by:

1. Quantifying the L/D enantiomer ratio generated by the O(3)-equivariant Boltzmann Generator for AD2
2. Probing the chiral stability and prior/latent space organization of the O(3) model via perturbation and clustering analysis

METHODS

Ascertaining Chirality of the Generated Molecule

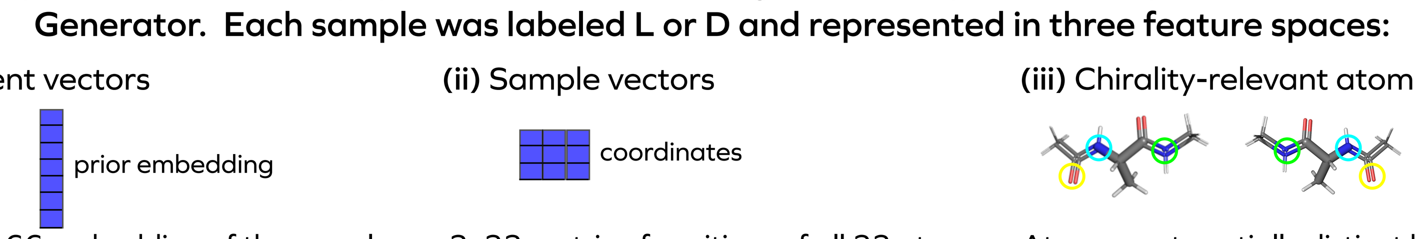


Generated AD2 molecule | Get CA,CB,N,C coordinates | Generate vectors from CA compute normal $\vec{n} = \vec{v}_C \times \vec{v}_N$ | Compute sign of dot product to ascertain chirality

Understanding Latent Space Organization

Analyzed 400 alanine dipeptide (AD2) structures generated by an O(3)-equivariant Boltzmann Generator. Each sample was labeled L or D and represented in three feature spaces:

(i) Latent vectors | (ii) Sample vectors | (iii) Chirality-relevant atom subset



Prior 1x66 embedding of the sample coordinates, which will be transformed through an O(3)-equivariant model.

3x22 matrix of positions of all 22 atoms in the molecule (66D vector). Direct encoding of raw spatial geometry.

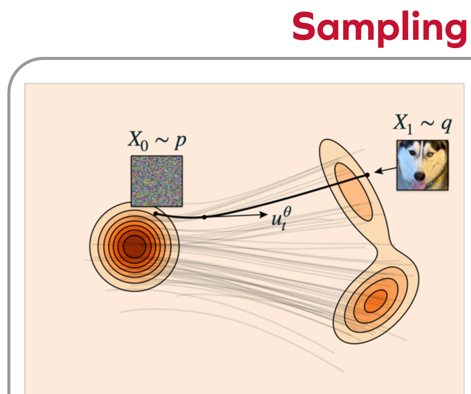
Atoms most spatially distinct between L- and D-enantiomers, used to isolate chirality signal.

Chirality-relevant atom selection equation:

$$\Delta_i = \| \mu_i^{(L)} - \mu_i^{D} \|_2$$

$$\text{Select atoms with largest per-atom difference: } \Delta_i > \tau$$

Sampling and Generating Perturbed Priors/Latent Samples



$$\frac{\mathrm{d}}{\mathrm{d}t} X_t = u_t^\theta(X_0)$$

$$\Longrightarrow X_1 = X_0 + \int_0^1 u_t^\theta(X_0)dt$$

To generate a conformer from the target distribution, we use **numerical ODE integrator 'dopri5'**

Figure Credit - Flow Matching Tutorial, Lipman et al., NeurIPS 2024

$$s_1 + \epsilon \times s_2 = \text{Perturbed Sample}$$

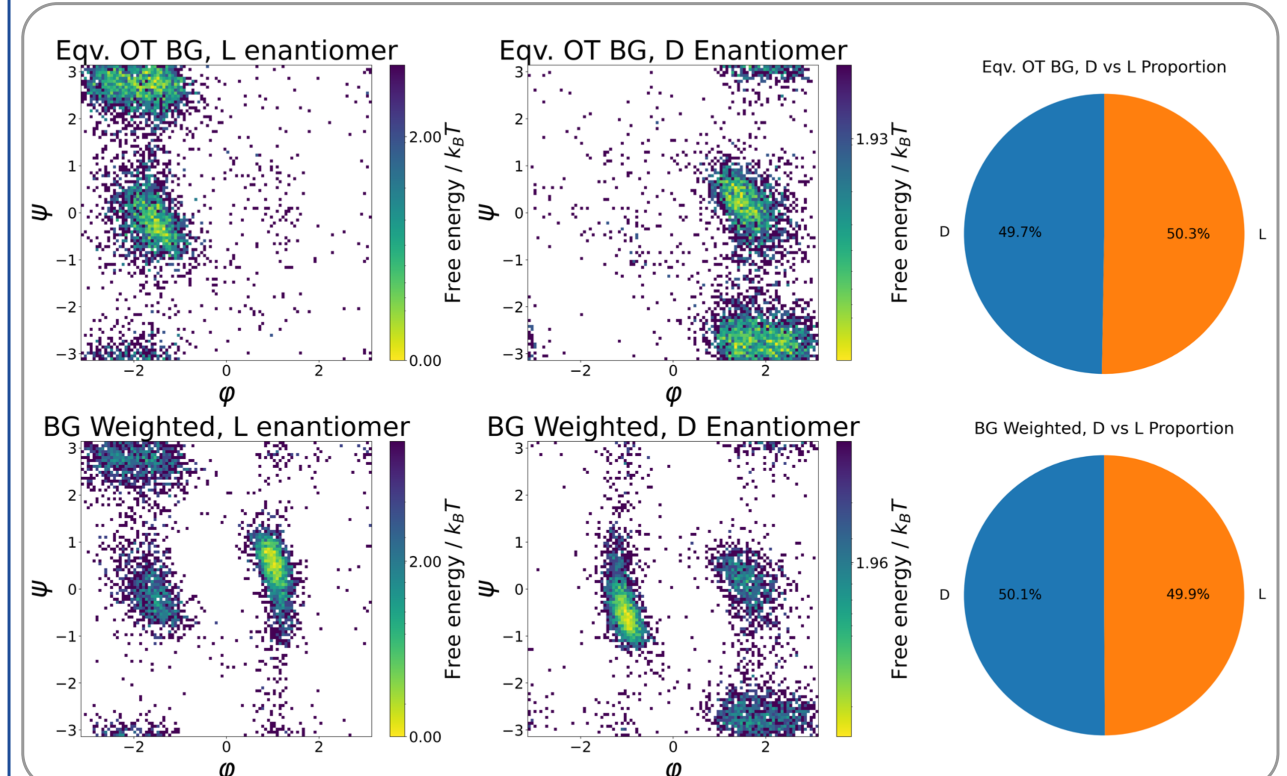
$s_1, s_2 \sim \text{Mean} - \text{Free } N(0, \Sigma) \text{ (Prior)}$

$\epsilon = \text{scaling parameter}$

This is the scheme we used to generate perturbations of varying strength

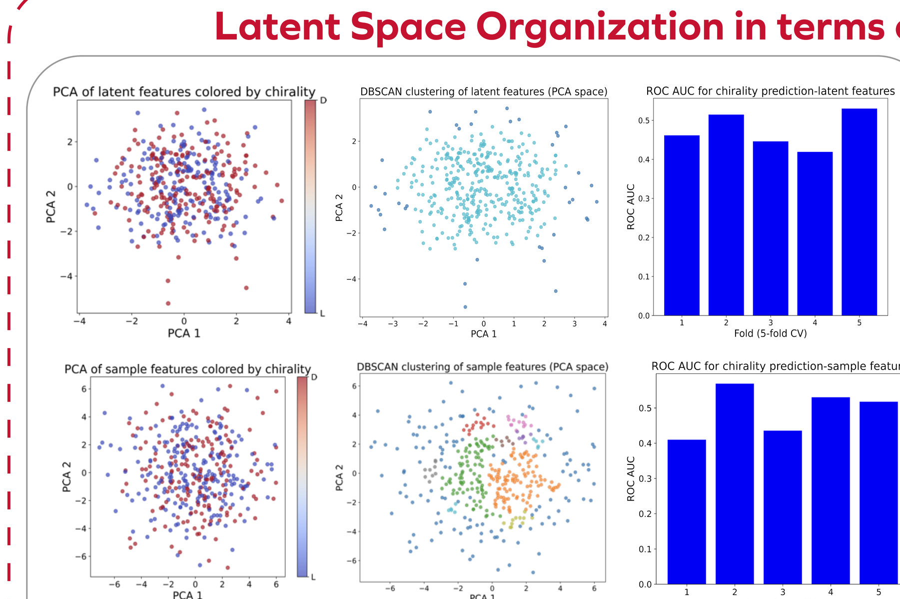
RESULTS

Quantifying Sample L/D Enantiomeric Proportions for weighted OT BG v/s Equiv. OT BG

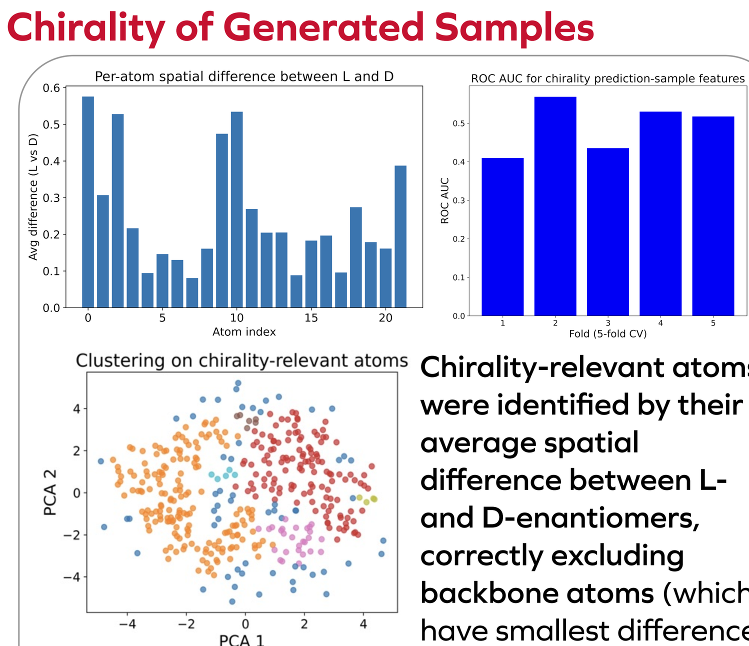


1. For both trained regimes (ie usage of OT cost vs equivariant OT cost), 10,000 conformations were generated. The proportion of L- and D-enantiomers was roughly equal (as expected)
2. This finding is corroborated by the Ramachandran plots.
3. While the proportion of enantiomers for both the models is roughly equal, we see a clear difference in the subset of conformers which can be attributed either to reweighting of the samples before training or the training regime.

Latent Space Organization in terms of Chirality of Generated Samples

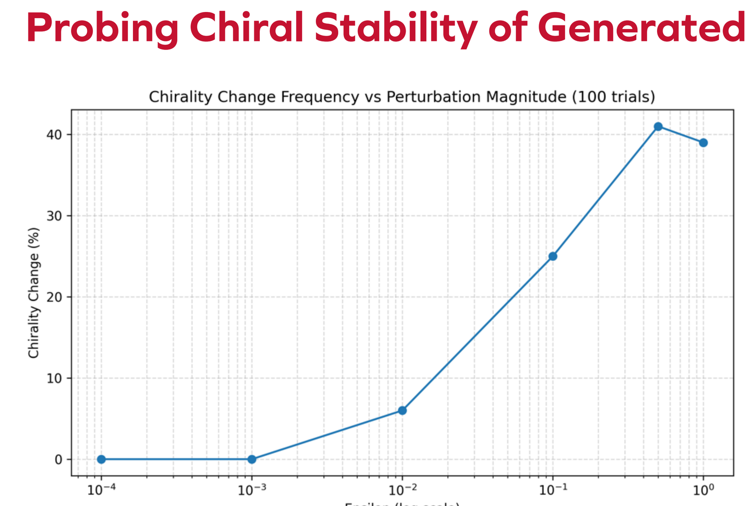


Clustering and classification based on latent and sample features shows limited separation of L- vs D-enantiomers. PCA followed by DBSCAN does not yield chirality-enriched clusters. Logistic regression ROC-AUC scores near 0.5 indicate that chirality is not linearly separable or structured within the clusters.

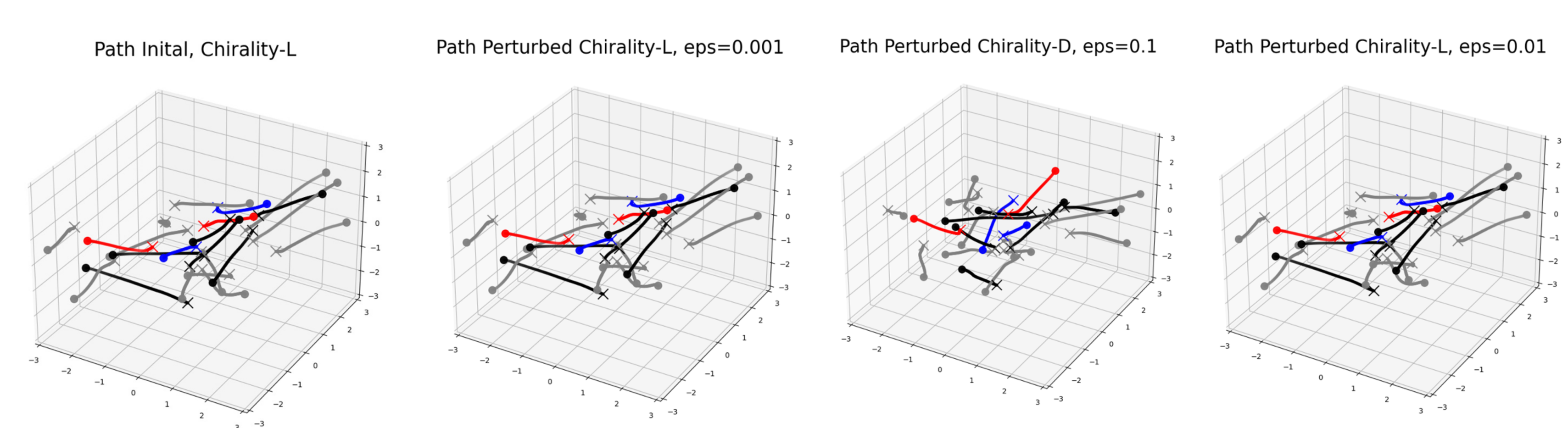


Chirality-relevant atoms were identified by their average spatial difference between L- and D-enantiomers, correctly excluding backbone atoms (which have smallest differences between enantiomers). Clustering improves using only chirality-relevant atoms, while logistic regression on full coordinates yields near-random performance, indicating chirality is weakly represented in the sample feature space.

Probing Chiral Stability of Generated Samples via Perturbations of the Latent Samples



1. Increasing the magnitude of latent space perturbation ***significantly increases*** the probability of the generating the complementary enantiomer.
2. Eps>0.001 was found to be the breaking point for chiral stability of generated samples.
3. Flow paths for the inital sample and perturbed samples are shown in the bottom panel. (X-latent,•-target,colors-atom type)



DISCUSSION

i. E(3)-equivariant BGs generate either an L-or D-enantiomer with roughly equal probability, however the training regime biases which conformers are sampled by the models.

ii. Clustering reveals absence of underlying separability of both prior/latent and sample features based only on chirality. This observation holds also when clustering only on chirality-relevant atoms.

iii. When generating biological molecules, (where enantiomeric purity is relevant), having only SO(3) rotation invariance instead of O(3) reflection and rotation equivariance is paramount.

iv. SE(3)-equivariant models instead of E(3)-equivariant models should be used when distinction between enantiomers is necessary.

v. In the future, we could train the E(3)-equivariant BG with trajectories from both the L- and D-enantiomers of the AD2 system individually and verify our hypothesis that the generated samples will retain the same L/D ratio.

REFERENCES

1. Frank Noé et al. ,Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. Science365,eaaw1147(2019).DOI:10.1126/science.aaw1147
2. Klein, L., Krämer, A., & Noé, F. (2023). Equivariant flow matching. Advances in Neural Information Processing Systems, 36, 59886–59910
3. Satorras, V. G., Hoogeboom, E., Fuchs, F. B., Posner, I., & Welling, M. (2021). E(n) Equivariant Normalizing Flows. Advances in Neural Information Processing Systems, 34, 4181–4192.
4. Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., & Le, M. (2022). Flow Matching for Generative Modeling. arXiv preprint arXiv:2210.02747.