Molecular Geometry Pretraining with SE(3)-Invariant Denoising Distance Matching

ICLR 2023

Shengchao Liu, *Hongyu Guo, Jian Tang*

Problem Definition

Pure 3D geometric setting.

- Pretraining: a large molecule 3D dataset (1M from Molecule3D [1]).
- Downstream tasks:
  - QM9: quantum mechanics prediction.
  - MD17: force prediction.
  - LBA & LEP: ligand-pocket binding prediction.

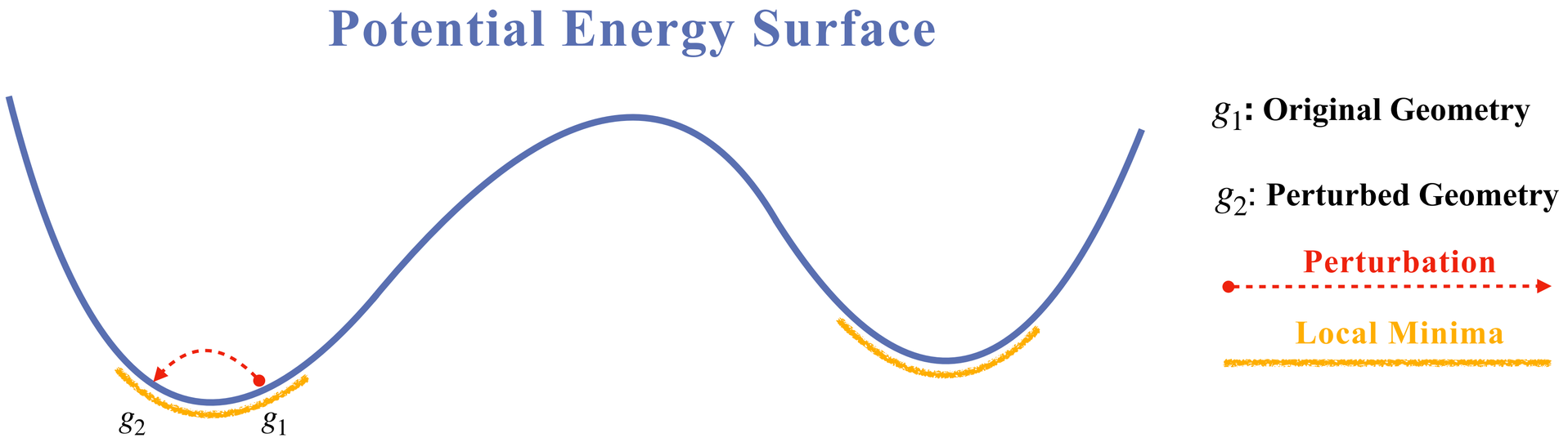
[1] Xu, Zhao, et al. "Molecule3D: A Benchmark for Predicting 3D Geometries from Molecular Graphs." *arXiv preprint arXiv:2110.01717* (2021).

Key Concepts: PES & Coordinate Perturbation

Coordinate perturbation is **important**!

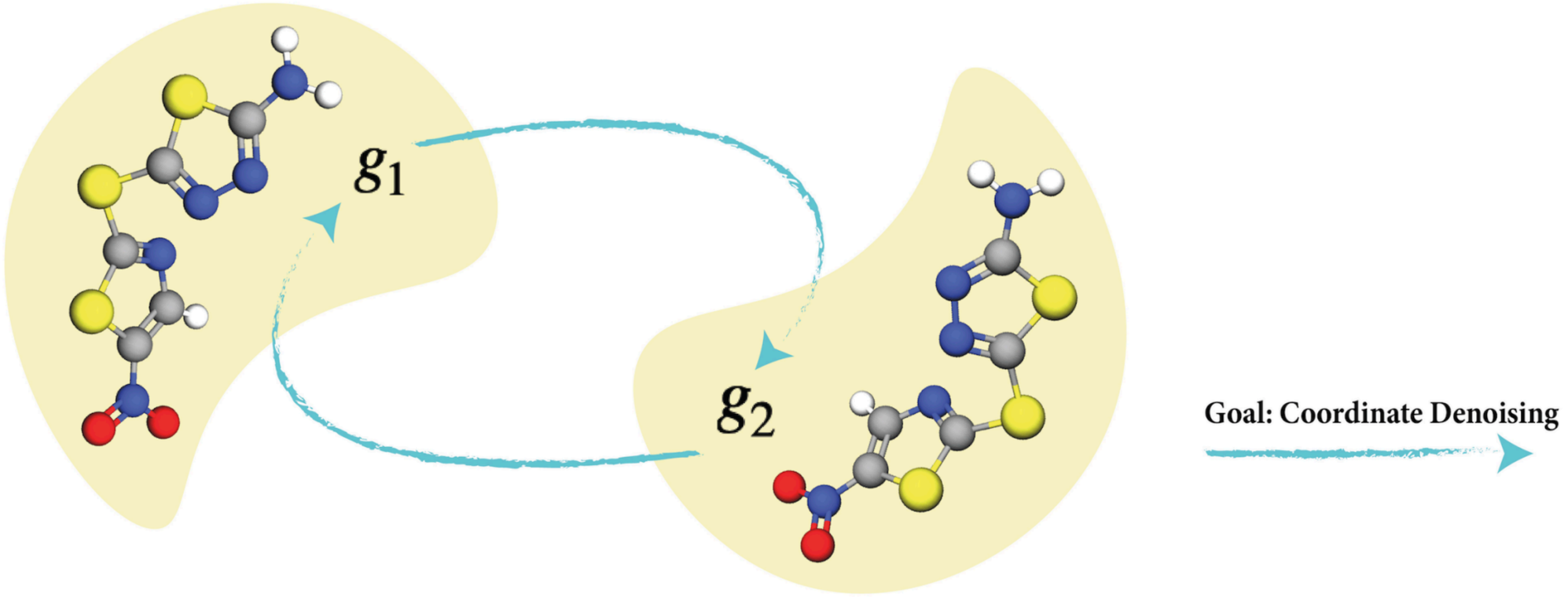
Table 5: An evidence example on molecular data. The goal is to predict 12 quantum properties (regression tasks) of 3D molecules (with 3D coordinates on each atom). The evaluation metric is MAE.

Model	Mode	Alpha ↓	Gap ↓	HOMO↓	LUMO ↓	Mu ↓	Cv ↓	G298 ↓	H298 ↓	R2 ↓	U298 ↓	U0 ↓	Zpve ↓
SchNet	Stable Geometry	0.070	50.59	32.53	26.33	0.029	0.032	14.68	14.85	0.122	14.70	14.44	1.698
	Type Corruption	0.074	52.07	33.64	26.75	0.032	0.032	21.68	22.93	0.231	23.01	22.99	1.677
	Coordinate Corruption	0.265	110.59	79.92	78.59	0.422	0.113	57.07	58.92	18.649	60.71	59.32	5.151
PaiNN	Stable Geometry	0.048	44.50	26.00	21.11	0.016	0.025	8.31	7.67	0.132	7.77	7.89	1.322
	Type Corruption	0.057	45.61	27.22	22.16	0.016	0.025	11.48	11.60	0.181	11.15	10.89	1.339
	Coordinate Corruption	0.223	108.31	73.43	72.35	0.391	0.095	48.40	51.82	16.828	51.43	48.95	4.395



GeoSSL: Geometric Self-supervised Learning

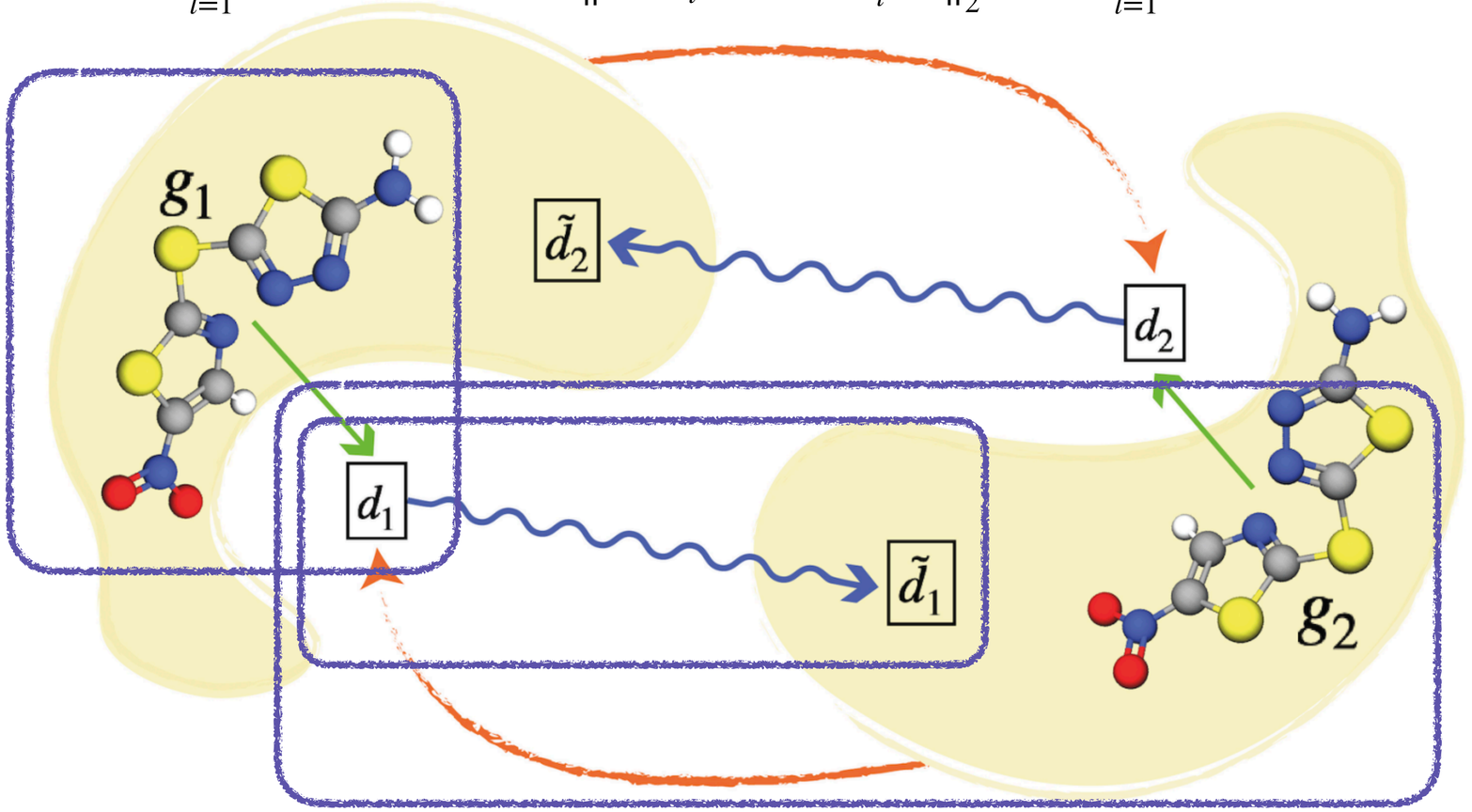
$$\mathscr{L}_{GeoSSL} = \frac{1}{2}\mathbb{E}_{p(g_1,g_2)}\left[\log p(g_1 \mid g_2)\right] + \frac{1}{2}\mathbb{E}_{p(g_1,g_2)}\left[\log p(g_2 \mid g_1)\right]$$



GeoSSL-DDM: Denoising Distance Matching

$$\mathscr{L}_{GeoSSL} = \frac{1}{2}\mathbb{E}_{p(d_1,g_2)}\Big[\log p(d_1 \mid g_2)\Big] + \frac{1}{2}\mathbb{E}_{p(g_1,d_2)}\Big[\log p(d_2 \mid g_1)\Big]$$

$$= \frac{1}{2L}\sum_{l=1}^{L}\sigma_l^{\beta}\mathbb{E}_{p(d_1|g_2)}\mathbb{E}_{q(\tilde{d}_1|d_1,g_2)}\left[\left\|\frac{s_\theta(\tilde{d}_1, g_2)}{\sigma_l} - \frac{d_1 - \tilde{d}_1}{\sigma_l^2}\right\|_2^2\right] + \frac{1}{2L}\sum_{l=1}^{L}\sigma_l^{\beta}\mathbb{E}_{p(d_2|g_1)}\mathbb{E}_{q(\tilde{d}_2|d_2,g_1)}\left[\left\|\frac{s_\theta(\tilde{d}_2, g_1)}{\sigma_l} - \frac{d_2 - \tilde{d}_2}{\sigma_l^2}\right\|_2^2\right]$$



Step 1: Distance Extraction

Step 2: Distance Perturbation

Step 3: Distance Denoising

Experiments

Table 1: Downstream results on 12 quantum mechanics prediction tasks from QM9. We take 110K for training, 10K for validation, and 11K for test. The evaluation is mean absolute error, and the best results are in **bold**.

Pretraining	Alpha ↓	Gap ↓	HOMO↓	LUMO ↓	Mu ↓	Cv ↓	G298 ↓	H298 ↓	R2 ↓	U298 ↓	U0 ↓	Zpve ↓
–	0.048	44.50	26.00	21.11	0.016	0.025	8.31	7.67	0.132	7.77	7.89	1.322
Supervised	0.049	45.33	26.61	21.77	0.016	0.026	8.97	8.59	0.170	8.35	8.19	1.346
Type Prediction	0.050	47.28	30.56	23.18	0.016	**0.024**	9.32	9.10	0.163	8.94	8.60	1.357
Distance Prediction	0.063	47.62	29.18	22.40	0.019	0.045	12.02	12.31	0.636	11.76	12.22	1.840
Angle Prediction	0.056	47.36	29.53	22.61	0.018	0.027	10.23	10.13	0.143	9.95	9.70	1.643
3D InfoGraph	0.053	44.79	27.09	21.66	0.016	0.027	9.22	8.78	0.143	8.94	9.11	1.465
GeoSSL-RR	0.048	44.85	25.42	20.82	**0.015**	0.025	8.56	8.20	0.133	7.89	7.62	1.329
GeoSSL-InfoNCE	0.052	45.65	26.70	21.87	0.016	0.027	9.17	9.62	0.130	8.77	8.63	1.519
GeoSSL-EBM-NCE	0.049	44.18	26.29	21.46	**0.015**	0.026	8.56	8.13	0.126	8.01	7.96	1.447
GeoSSL-DDM (ours)	**0.046**	**40.22**	**23.48**	**19.42**	**0.015**	**0.024**	**7.65**	**7.09**	**0.122**	**6.99**	**6.92**	**1.307**

Table 2: Downstream results on 8 force prediction tasks from MD17. We take 1K for training, 1K for validation, and the number of molecules for test are varied among different tasks, ranging from 48K to 991K. The evaluation is mean absolute error, and the best results are in **bold**.

Pretraining	Aspirin ↓	Benzene ↓	Ethanol ↓	Malonaldehyde ↓	Naphthalene ↓	Salicylic ↓	Toluene ↓	Uracil ↓
–	0.556	0.052	0.213	0.338	0.138	0.288	0.155	0.194
Supervised	0.478	0.145	0.318	0.434	0.460	0.527	0.251	0.404
Type Prediction	1.656	0.349	0.414	0.886	1.684	1.807	0.660	1.020
Distance Prediction	1.434	0.090	0.378	1.017	0.631	1.569	0.350	0.415
Angle Prediction	0.839	0.105	0.337	0.517	0.772	0.931	0.274	0.676
3D InfoGraph	0.844	0.114	0.344	0.741	1.062	0.945	0.373	0.812
GeoSSL-RR	0.502	0.052	0.219	0.334	0.130	0.312	0.152	0.192
GeoSSL-InfoNCE	0.881	0.066	0.275	0.550	0.356	0.607	0.186	0.559
GeoSSL-EBM-NCE	0.598	0.073	0.237	0.518	0.246	0.416	0.178	0.475
GeoSSL-DDM (ours)	**0.453**	**0.051**	**0.166**	**0.288**	**0.129**	**0.266**	**0.122**	**0.183**

Thank you!