MGB: The Material Generation Benchmark

Liang Yan [1,2] Beom Seok Kang [2] Maurice D. Hanisch [2] Jian Ma [3] Anima Anandkumar [2]

[1]Fudan University [2]Caltech [3]Shanghai Jiao Tong University

The Material Discovery Pipeline

Generative models learn the relationship between structure and properties and, under chemistry and symmetry constraints, sample latent space to generate high quality candidates. After standardization, deduplication, and quick surrogate screening, only the best designs undergo costly DFT and MD calculations. Validation feedback refines the model, **forming a continuous loop that reduces search space, improves hit rates, and lowers computation, enabling faster discovery of materials such as inorganic crystals and MOFs.**

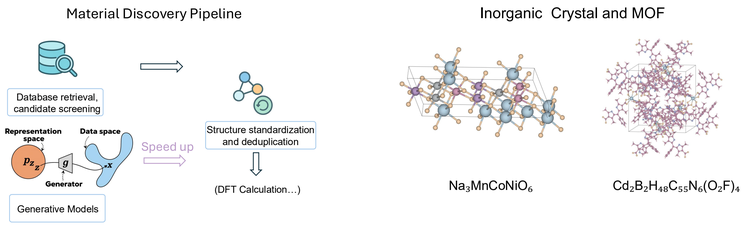
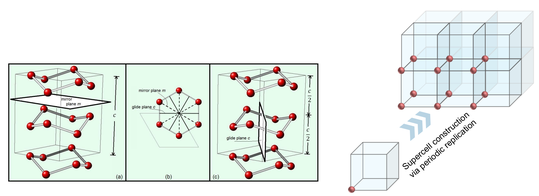


Figure 1. The material discovery pipeline accelerated by generative models.

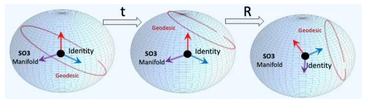
Physical Properties of Crystals

Crystal properties arise from three key principles: **(1)** Space Group: space group operations such as mirrors, rotation axes, and glide planes define allowed atomic arrangements **(2)** Periodicity: Repetition of one unit cell through 3D translations forms the full lattice. **(3)** Geometric modeling: Representing translations and rotations in SE(3) provides a compact, symmetry-aware framework.



(1) Space Group of Crystal.

(2) Periodicity of Crystal Structure.



(3) Visualization of SE3 group using a Hypersphere for translation and rotations.

Figure 2. Illustration of physical properties of crystals. (1) The space group of hexagonal H2O ice is P63/mmc. (a) The first m indicates the mirror plane perpendicular to the c-axis, (b) the second m indicates the mirror planes parallel to the c-axis, and (c) the c indicates the glide planes (b) and (c). The black boxes outline the unit cell. (2) Illustration of the periodicity of crystal structures. (3) Visualization of SE3 group using a Hypersphere for translation and rotations.

MGB: The Material Generation Benchmark

Tasks.

(1) **Crystal Structure Prediction (CSP)**, $p(L, F \mid A)$; (2) **De Novo Generation (DNG)**, $p(\mathcal{M})$; (3) **MOF Structure Prediction (MSP)**; and (4) **Out-of-Distribution (OOD) Generation**.

Models.

(1) **Diffusion/flow methods**: DiffCSP, DiffCSP++, EquiCSP, FlowMM, MOFFlow;
(2) **Hybrid models**: CDVAE, Cond-CDVAE;
(3) **Large Language Models**: CrystalLLM-25M/200M, Llama 3.1 (8B).

Datasets.

(1) **Carbon-24**: a dataset of carbon crystal structures containing fewer than 24 atoms per unit cell.
(2) **Perov-5**: a collection of perovskite structures following the ABX_3 formula.
(3) **MP-20**: a subsets of inorganic crystal structures obtained from the Materials Project.
(4) **MPTS-52**: a subsets of inorganic crystal structures obtained from the Materials Project.
(5) **Boyd MOF**: A set of hypothetical metal–organic frameworks curated for evaluating CO_2/N_2 separation performance.

Evaluation.

(1) **Matching Accuracy**: match rate and coordinate RMSE.
(2) **Generation Quality**: validity, diversity, and distribution alignment.
(3) **OOD Generation**: novel, valid samples under deliberate shifts.
(4) **Physical Plausibility**: periodic atomic-collision checks and ratios.
(5) **Symmetry Awareness**: translation invariance via the IPT metric.
(6) **Computational Complexity**: model size and inference time (efficiency/scalability).

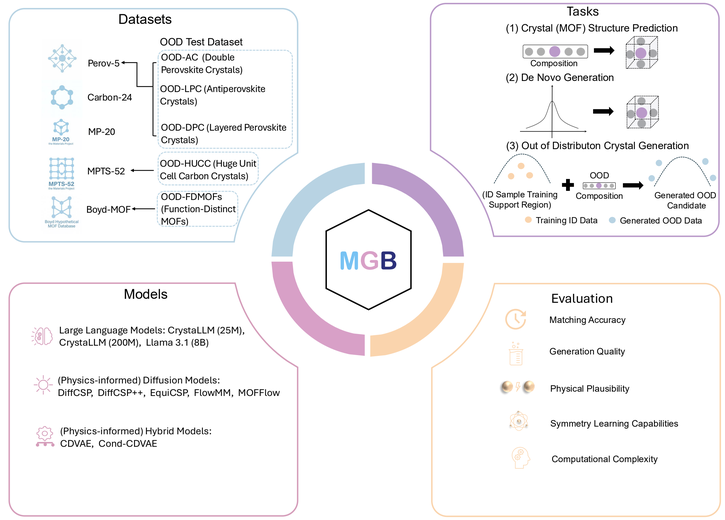


Figure 3. The MGB pipeline, including datasets, model families, benchmark tasks (structure prediction, de novo generation, and OOD generation), and evaluation metrics.

The Benchmarking Results of Crystal Structure Prediction

Table 1. The benchmarking results on the crystal structure prediction task for diffusion-based models.

Method	# of samples	Perov-5		Carbon-24		MP-20		MPTS-52	
		MR (↑)	RMSE (↓)	MR (↑)	RMSE (↓)	MR (↑)	RMSE (↓)	MR (↑)	RMSE (↓)
Cond-CDVAE	1	42.31	0.1356	14.65	0.3216	29.91	0.1098	4.91	0.2387
DiffCSP	1	51.81	0.0922	16.45	0.2865	47.07	0.0654	11.91	0.1493
DiffCSP++ (w/ CSPML)	1	53.71	0.0880	-	-	70.94	0.0295	33.17	**0.0893**
DiffCSP++ (w/ GT)	1	**98.47**	**0.0398**	-	-	**79.76**	0.0293	42.13	0.1134
EquiCSP	1	51.89	0.0746	17.19	0.2751	52.33	0.0612	13.04	0.1293
FlowMM	1	47.38	0.1183	15.53	0.2848	50.21	0.1192	8.20	0.2275
CrystaLLM-raw$_{(25M)}$	1	47.95	0.0966	21.13	**0.1687**	55.85	0.0437	17.47	0.1113
CrystaLLM$_{(25M)}$	1	45.65	0.0977	**21.87**	0.1734	56.58	0.0426	17.54	0.1028
CrystaLLM-raw$_{(200M)}$	1	46.10	0.0953	20.25	0.1761	58.70	0.0408	19.21	0.1110
CrystaLLM$_{(200M)}$	1	45.87	0.0970	20.64	0.1971	58.98	**0.0345**	18.97	0.1123

The Benchmarking Results of De Novo Generation

Table 2. The benchmarking results on de novo generation task (N=10000) with best checkpoint.

Dataset	Method	Validity (↑)		Coverage (↑)		Property (↓)		AMSD (↓)		AMCD (↓)	
		Comp.	Struc.	COV-R	COV-P	d_ρ	d_{elem}	AMSD-R	AMSD-P	AMCD-R	AMCD-P
Perov-5	CDVAE$_{(b)}$	**98.69**	100.00	99.15	98.58	0.2249	0.1569	0.0476	0.0561	0.7356	1.262
	Cond-CDVAE$_{(b)}$	98.55	100.00	99.29	98.51	0.0786	0.1279	0.0485	0.0603	0.7202	**1.244**
	DiffCSP$_{(b)}$	98.66	100.00	**99.66**	98.25	0.1370	0.0542	0.0383	0.0437	**0.6360**	1.280
	DiffCSP++$_{(b)}$	98.49	100.00	**99.66**	**98.74**	**0.0607**	**0.0369**	**0.0265**	**0.0276**	0.6491	1.290
	EquiCSP$_{(b)}$	97.84	100.00	99.34	98.44	0.1200	0.0569	0.0461	0.0379	0.6698	1.298
	FlowMM$_{(b)}$	98.24	99.99	99.31	89.36	0.6881	0.0730	0.0455	0.0803	0.6745	1.367
Carbon-24	CDVAE$_{(b)}$	100.00	99.99	99.85	85.95	**0.1093**	-	0.0438	0.1230	0.00	0.00
	Cond-CDVAE$_{(b)}$	100.00	100.00	99.70	73.93	0.2718	-	0.0547	0.1651	0.00	0.00
	DiffCSP$_{(b)}$	100.00	100.00	99.90	93.61	0.1429	-	**0.0190**	0.0523	0.00	0.00
	DiffCSP++$_{(b)}$	100.00	99.98	99.56	46.02	0.1732	-	0.0370	0.3437	0.00	0.00
	EquiCSP$_{(b)}$	100.00	100.00	99.90	94.97	0.1501	-	0.0221	0.0511	0.00	0.00
	FlowMM$_{(b)}$	100.00	99.94	**100.00**	**95.48**	0.1684	-	0.0257	**0.0443**	0.00	0.00
MP-20	CDVAE$_{(b)}$	**86.14**	99.97	99.39	99.39	0.6412	1.269	0.1554	0.1906	3.538	3.996
	Cond-CDVAE$_{(b)}$	85.68	**99.99**	98.84	99.35	0.4805	1.426	0.1660	0.1923	3.671	4.256
	DiffCSP$_{(b)}$	83.96	99.78	99.61	99.47	**0.1027**	0.6130	0.1079	**0.1271**	2.996	3.475
	DiffCSP++$_{(b)}$	84.38	99.80	**99.73**	99.34	0.1996	0.4548	**0.1040**	0.1349	2.976	3.510
	EquiCSP$_{(b)}$	81.16	99.92	99.51	**99.61**	0.1357	0.4728	0.1201	0.1360	3.045	3.524
	FlowMM$_{(b)}$	83.66	92.26	99.19	99.24	1.070	**0.1189**	0.1407	0.1663	**2.660**	**2.794**
MPTS-52	CDVAE$_{(b)}$	81.17	99.98	99.69	98.07	**0.5166**	1.032	0.1699	0.1764	3.737	4.555
	Cond-CDVAE$_{(b)}$	**82.70**	**100.00**	99.69	**98.69**	0.1268	0.9046	0.1709	0.1740	3.705	4.484
	DiffCSP$_{(b)}$	66.70	99.78	99.64	88.89	0.9409	0.5573	0.1758	0.1596	3.765	4.600
	DiffCSP++$_{(b)}$	78.39	99.84	**99.86**	96.97	1.051	**0.4053**	**0.1682**	0.1727	3.497	4.631
	EquiCSP$_{(b)}$	66.56	99.66	99.72	96.80	0.8208	0.5493	0.1785	**0.1584**	3.623	4.243
	FlowMM$_{(b)}$	72.73	93.92	99.53	98.28	2.134	0.4779	0.1857	0.1695	**3.395**	**4.175**

The Evaluations on the Physical Plausibility Problem

Table 3. The atomic collision benchmarking results on crystal structure prediction task (global seed).

Dataset	Method	# Crystals	Collided (↓)	Collision Rate (↓)	# Collision Pairs (↓)	Cross-cell (↓)	Same-cell (↓)
Perov-5	DiffCSP	3785	350	9.25%	769	364 (47.33%)	405 (52.67%)
	DiffCSP++ (w/ GT)	3785	375	9.91%	860	198 (23.02%)	662 (76.98%)
	DiffCSP++(w/ CSPML)	3785	376	9.93%	944	251 (25.59%)	693 (73.41%)
	EquiCSP	3785	367	9.70%	835	444 (53.17%)	391 (46.83%)
MP-20	DiffCSP	9046	945	10.45%	2912	1434 (49.24%)	1478 (50.76%)
	DiffCSP++ (w/ GT)	9046	664	7.34%	4848	2103 (43.38%)	2745 (56.62%)
	DiffCSP++ (w/ CSPML)	9046	669	7.40%	9103	4089 (44.92%)	5014 (55.08%)
	EquiCSP	9046	1079	11.93%	3036	1494 (48.73%)	1572 (51.27%)
MPTS-52	DiffCSP	8096	3008	37.15%	13818	4426 (32.03%)	9392 (67.97%)
	EquiCSP	8096	2875	35.51%	12138	3628 (29.89%)	8510 (70.11%)