Open Molecular Crystals 2025 (OMC25) Dataset and Models

 $\label{eq:local_problem} \begin{subarray}{ll} Vahe Gharakhanyan^1, Luis Barroso-Luque^1, Yi Yang^2, Muhammed Shuaibi^1, Kyle Michel^1, Daniel S. \\ Levine^1, Misko Dzamba^1, Xiang Fu^1, Meng Gao^1, Xingyu Liu^3, Haoran Ni^2, Keian Noori^3, Brandon M. \\ Wood^1, Matt Uyttendaele^1, Arman Boromand^3, C. Lawrence Zitnick^1, Noa Marom^{2,4,5}, Zachary W. Ulissi^1, Anuroop Sriram^1\\ \end{subarray}$

¹Fundamental AI Research at Meta, ²Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA, USA, ³Reality Labs Research at Meta, ⁴Department of Physics, Carnegie Mellon University, Pittsburgh, PA, USA, ⁵Department of Chemistry, Carnegie Mellon University, Pittsburgh, PA, USA

The development of accurate and efficient machine learning models for predicting the structure and properties of molecular crystals has been hindered by the scarcity of publicly available datasets of structures with property labels. To address this challenge, we introduce the Open Molecular Crystals 2025 (OMC25) dataset, a collection of over 27 million molecular crystal structures containing 12 elements and up to 300 atoms in the unit cell. The dataset was generated from dispersion-inclusive density functional theory (DFT) relaxation trajectories of over 230,000 randomly generated molecular crystal structures of around 50,000 organic molecules. OMC25 comprises diverse chemical compounds capable of forming different intermolecular interactions and a wide range of crystal packing motifs. We provide detailed information on the dataset's construction, composition, structure, and properties. To demonstrate the quality and use cases of OMC25, we further trained and evaluated state-of-the-art open-source machine learning interatomic potentials. By making this dataset publicly available, we aim to accelerate the development of more accurate and efficient machine learning models for molecular crystals.

Dataset: https://huggingface.co/facebook/OMC25

Models: https://huggingface.co/facebook/OMC25 and https://huggingface.co/facebook/UMA

Code: https://github.com/facebookresearch/fairchem

Correspondence: V.G. (vaheg@meta.com), A.S. (anuroops@meta.com), Z.W.U. (zulissi@meta.com), N.M.

(nmarom@andrew.cmu.edu)

Meta

Background & Summary

Molecular crystals are a class of materials characterized by the orderly arrangement of molecules in a crystalline lattice. These materials have important applications in pharmaceuticals [1–5], organic electronics [6–9], and other fields [10–15], due to their unique structural and functional properties. A key phenomenon in molecular crystals is polymorphism, where a single molecule can form multiple crystal structures, influencing the physical properties of the material [16, 17]. Understanding, predicting, and controlling the formation of different polymorphs is crucial for optimizing the properties of molecular crystals for tailored applications [16]. This requires exploring the potential energy surfaces of different crystal structures to gain insight into their relative stability.

Computer simulations have become indispensable tools in the study of molecular crystals. In recent years, the advent of machine learning (ML) has revolutionized the fields of chemistry and materials science [18–29]. Machine learning interatomic potentials (MLIPs) trained on large *ab initio* datasets offer a promising alternative to density functional theory (DFT), achieving similar accuracy at a fraction of the computational cost. The accuracy and computational demands of MLIPs typically range between those of classical force fields and DFT, depending on the model complexity and the training domain [30]. This middle ground enables MLIPs to strike an effective balance between computational efficiency and accuracy, making them

Open Molecular Crystals 2025

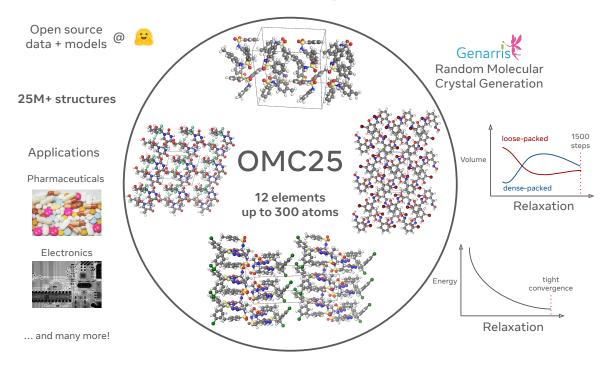


Figure 1 Overview of the OMC25 dataset: generation method, structure relaxations, statistics, and application areas.

well-suited for large-scale simulations, such as crystal structure prediction (CSP), where using DFT would be too costly. The efficacy of ML models, however, is contingent upon the quality and diversity of the training data. The development of novel MLIP architectures has been facilitated by the release of large, diverse, and open-source datasets tailored to molecules (OMol25 [31], QM9 [32], and others [33–37]) and inorganic materials (OMat24 [38], OC20 [39], and others [40–42]) applications. MLIPs used for molecular crystals have been trained predominantly on data for isolated molecules [43–47]. Some have been trained on small-scale [48, 49] or proprietary [50] molecular crystal datasets. A significant gap remains in the availability of large-scale open datasets specifically designed for molecular crystal applications. This limitation hinders the advancement of MLIPs in this domain, underscoring the need for a comprehensive dataset that can provide a rich source of structural and property information for training machine learning models.

We present the Open Molecular Crystals 2025 (OMC25) dataset, a large-scale resource for training MLIPs for molecular crystals. OMC25 comprises over 27 million molecular crystal structures, containing 12 elements and up to 300 atoms in the unit cell. Each structure is labeled with total energy, atomic forces, and unit cell stress values. The structures comprising OMC25 were extracted from the dispersion-inclusive DFT relaxation trajectories of over 230 thousand putative molecular crystals constructed from 50 thousand unique molecules from the OE62 dataset [51]. The DFT data was acquired using the Perdew-Burke-Ernzerhof (PBE) [52] exchange-correlation functional combined with the Grimme D3 [53] dispersion correction (PBE-D3) with tight convergence settings. Diverse sampling of molecular packing arrangements across space groups with a varying number of molecules per unit cell (Z) was achieved using the open-source crystal generation software Genarris 3.0 [54]. To thoroughly sample the potential energy landscape, including regions far from equilibrium, we generated both loosely packed and densely packed structures. To demonstrate the usefulness of the OMC25 dataset, we train MLIPs and evaluate their performance on established community benchmarks.

The dataset, model checkpoints, and code to train and evaluate models are all released open source to ensure reproducibility and to allow the community to build upon and further improve our results. We provide the OMC25 dataset with a CC BY 4.0 license and model weights with a commercially permissive license (with

some geographic and acceptable use restrictions). By making this dataset openly available, we aim to catalyze further research and enable advancements in structure and property prediction of molecular crystals.

Methods

As machine learning interatomic potentials (MLIPs) see increasing adoption for use in molecular crystal research, an open, comprehensive dataset specifically designed to cover molecular crystals has become an urgent need. To achieve this, we employed a multi-step process to curate, pre-process, label, and validate a diverse set of molecular crystal structures, covering various chemical compositions, crystal systems, and space groups. The resulting OMC25 dataset includes a wide range of property-labeled molecular crystal structures, reflecting the rich diversity of molecular crystals found in nature and synthetic materials, and is designed to be a valuable resource for advancing materials research.

Sampling Molecules. The OE62 dataset [51] served as our starting point for sampling molecular structures. The OE62 dataset includes molecules extracted from the Cambridge Structural Database (CSD) [55] repository of experimentally determined molecular crystal structures. The OE62 dataset contains 61,489 molecules comprising the elements H, Li, B, C, N, O, F, Si, P, S, Cl, As, Se, Br, Te, and I, whose geometry was optimized with DFT through the FHI-aims all electron code [56–58] with the Perdew, Becke, and Ernzerhof (PBE) functional [52] and the Tkatchenko-Scheffler (TS) dispersion correction [59] until the residual forces on each atom were below 0.001 eV/Å. This dataset was sampled for molecular geometries. Potentially energetic molecules were removed, resulting in approximately 50 thousand unique molecules. Owing to the scarcity of distinct conformers for a given molecule in the OE62 dataset, only one molecular conformer was obtained in the vast majority of cases. This means that, within our data generation workflow, the molecular conformation could change only to the extent possible during final geometry relaxation in the crystal, as described below.

Molecular Crystal Structure Generation. Random molecular crystal generation was performed with the Genarris 3.0 package [54]. Genarris is an open-source software that generates diverse molecular crystal packing arrangements starting from the input molecular conformer and Z number, the number of molecules in the unit cell. For each sampled molecule from the OE62 dataset, we selected two values of Z out of the six most frequent Z numbers in the CSD ($Z \in [4, 2, 8, 1, 16, 6]$, in order of prevalence), with the probability distributions derived from the CSD [55]. Genarris automatically identifies all compatible space groups matching the point group symmetry of the input molecular conformer and Z number with one molecule in the asymmetric unit of the crystal (Z'=1) [60]. For the purpose of generating a diverse set of random structures, as opposed to performing crystal structure prediction, we did not aim to exhaustively sample the configuration space of structures within a given space group. Therefore, for each compatible space group, we generated only two structures.

In order to train MLIPs that can work well in all regions of the potential energy surface (PES), we took advantage of the features of Genarris 3.0 to generate both loosely packed and close-packed structures. Genarris samples unit cell volumes from a Gaussian distribution around a target volume, which is estimated using the integrated machine learning (ML) model from PYMOVE [61]. To generate loosely-packed initial structures, we scaled the target volume by a factor of 1.25. Genarris avoids generating unphysical structures by demanding that the interatomic distances, d_{ij} , between atoms i and j from different molecules are greater than a cutoff: $d_{ij} > s_r(r_i^{\text{vdW}} + r_j^{\text{vdW}})$, where $r_{i/j}^{\text{vdW}}$ are the van der Waals radii of atoms i and j, respectively, and s_r is a user-defined scaling factor. In the initial generation step, we used $s_r = 0.95$ to ensure adequate spacing between molecules. Then, the initial loosely packed structures were optimized with the $Rigid\ Press\ algorithm$, implemented in Genarris 3.0 [54] to achieve close packing. Rigid press freezes the molecular geometry and employs a regularized hard-sphere potential to optimize the molecular position and orientation along with the crystal lattice vectors to compress the unit cell as much as possible, while preserving the space group symmetry. For the Rigid Press optimization we applied the default values of $s_r = 0.85$ and specialized s_r values for hydrogen bonds, derived from the statistical analysis of crystal structures in the CSD to ensure that molecules are as close to each other as possible without overlap [60].

Sampling Random Molecular Crystal Structures. To maximize the diversity of the selected putative structures, we sampled both loosely packed structures from the initial generation stage and close-packed structures from the Rigid Press stage of Genarris. The maximum number of atoms in the unit cell was capped at 300. For

each compound and chosen Z number, we selected up to two structures (where possible) whose geometry was converged within 5,000 iterations in the Rigid Press stage. Additionally, we sampled up to two structures from the initial generation stage, excluding the structure identifiers already sampled from the Rigid Press stage. This procedure leads to up to four structures for a given Z number. As noted above, we selected two Z numbers per molecule, producing a maximum of 8 putative crystal structures for each molecular conformer. In practice, only 4.7 structures on average were sampled per molecule. This approach allowed us to maximize the diversity of putative structures while minimizing redundancy, resulting in around 230,000 sampled putative crystal structures from Genarris.

Structural Relaxations. The molecular crystal structures selected in the previous step were fully relaxed using dispersion-corrected DFT. The calculations were performed using the Vienna Ab initio Simulation Package (VASP) [62–64] with the projector augmented wave (PAW) pseudopotentials [65, 66]. The PBE generalized gradient approximation (GGA) [52] was combined with the Grimme D3 dispersion correction [53]. The atomic positions and lattice vectors were relaxed until the maximum per-atom residual forces were below 0.001 eV/Å, or the relaxation required more than 1,500 steps. The total energy convergence tolerance was set to 0.001 meV and the plane-wave energy cut-off was 520 eV. The k-point grids were automatically set by the PYMATGEN library [67]. A detailed description of additional VASP input flags is provided in the Supplementary Information. The relaxation of the molecular crystal structures sampled from Genarris resulted in more than 300 million ionic steps and 1.5 billion electronic self-consistency steps.

Relaxation Trajectory Filtering. Regardless of the structural relaxation convergence, all trajectories were retained. Initially, we removed frames with non-negative energies, residual forces exceeding 50 eV/Å, and stresses above 80 GPa from the trajectory. Subsequently, if the final volume or the volume immediately following the first step deviated from the initial volume by more than 33%, the entire trajectory was discarded. However, if only a specific frame's volume differed by more than 33% from the initial one, that frame alone was removed. This was done to make sure that the initially selected k-point density was still deemed sufficient for the ionic step. Rarely, we observed that in some cases molecular fragments broke apart or merged during relaxation. To detect this, the connectivity of each structure frame was represented as an undirected graph object in Network [68] from the StructureGraph object of Graphs module obtained using the Jmolnn class from the local_env module of Pymatgen.Analysis [67]. Different frames were compared with an exact graph isomorphism check as implemented in the isomorphism module of Network algorithms [68]. If the molecular connectivity after the first relaxation step differed from the initial connectivity, the entire trajectory was discarded. Otherwise, any frames with altered connectivity compared to the starting structure were removed. This was done to make sure that the structure optimization was not heading towards a non-realistic molecular crystal or leading to chemical reactions.

Structure Sampling. For the purpose of training MLIPs, it is important to sample different regions of the potential energy landscape, both around the local minima and far from them. Therefore, after filtering the relaxation trajectories, a sampling strategy was employed to select a representative subset of frames. The goal was to capture the most informative and diverse set of structures along the relaxation trajectories while minimizing redundancy. A subset of up to 100 structure frames was sampled from each of the remaining trajectories, with the goal of maximizing their absolute energy differences from the preceding structure frame. In addition to energy difference-based sampling, approximately 20 total structure frames were uniformly sampled between the first occurrences in the trajectory of structures with maximum per-atom residual forces of 0.1, 0.01, and 0.001 eV/Å. The sampling strategy resulted in the final dataset including around 120 frames per molecular crystal relaxation trajectory, on average including around 10% of all frames in the original trajectories. This led to a total sample size of 27 million structures from the relaxation trajectories.

Table 1 Size, starting molecular crystal and molecule count of different OMC25 dataset splits.

Split	Size	Molecular crystals	Molecules	Fraction			
Train Val Test	24,870,226 1,386,816 1,358,143	207,271 11,570 11,327	44,403 2,467 2,467	90% 5% 5%			
Total	27,615,185	230,168	49,337				

Dataset Splits. The sampling strategy described above yielded the final structures included in the OMC25

dataset. To train MLIPs, we created the OMC25 training, validation, and test splits. To prevent leakage and ensure data integrity, an allocation process was implemented, wherein all frames belonging to putative structures of the same compound were assigned to a single split exclusively. A 90/5/5 random sampling strategy was adopted, where 90% of the data points were assigned to the training set (Train), 5% were assigned to the validation set (Val), and 5% to the test set (Test). The dataset sizes and compositions for each split of OMC25 are presented in Table 1. MLIPs were trained and optimized using the training and validation subsets of the dataset, with their performance subsequently assessed on a held-out test set, as well as through additional evaluation tests and metrics.

Data Records

The training and validation splits of the OMC25 dataset and related files are available for download from HuggingFace at https://huggingface.co/facebook/OMC25, under the CC BY 4.0 license, after applying for the repository access on HuggingFace.

We prepared all data files in the Atomic Simulation Environment (ASE) Lightning Memory-Mapped Database (LMDB) database format. LMDB format [69] is an efficient dataset type for large scale data storage designed around the concept of key-value pairs. Users can read and query the datasets using the ASE DB API [70], where each entry (value) is an ASE.ATOMS object that includes information on lattice parameters, atomic positions and numbers, periodic boundary conditions, total structure energy, atomic forces, and unit cell stress. Additionally, for each entry, we also provide information on the Cambridge Structural Database (CSD) [55] reference code (csd_refcode) corresponding to the molecule from the OE62 dataset (taken directly from [51]), the Z value of the unit cell (z_value), genarris_step tag of either the generation (gener) or the Rigid Press (press) stage of Genarris 3.0 [54], xtal.id unique crystal identifier among putative structures from Genarris step, and sid structure identifier consisting of the above information and also including the index of the structure frame in the filtered relaxation trajectory. We also provide detailed information on all unique initial molecular crystal structures that underwent structural relaxations in the omc25-starting-crystals.csv file (Table 2).

Table 2 Description of columns in omc25-starting-crystals.csv describing all starting molecular crystal structures that underwent structural relaxations.

Column name	Description				
csd_refcode	CSD reference code [55] of molecule from OE62 dataset [51]				
z_value	Number of molecules in the crystal unit cell				
genarris_step	Sampled from generation (gener) or Rigid Press (press) step				
	of Genarris 3.0 [54]				
xtal.id	Unique crystal identifier among putative structures from				
Xtar.iu	Genarris step				
anlit	Structure was included in the training (train) or validation				
split	(val) split				
nframes	Number of frames sampled from relaxation trajectory				
mol.composition, xtal.composition	Composition of molecule and crystal, respectively				
	Number of atoms in molecule and crystal unit cell, respec-				
mol.natoms, xtal.natoms	tively				
mol maga wtal maga	Molar mass in g/mol of molecule and crystal unit cell, re-				
mol.mass, xtal.mass	spectively				
xtal.spacegroup	Crystal space group				

Technical Validation

A robust dataset for machine learning interatomic potentials (MLIPs) must comprehensively sample the relevant chemical, structural, and property spaces while maintaining high data quality. Here, we detail the

diversity and validation of the OMC25 dataset, highlighting both its breadth, the rigorous controls applied, and the performance of MLIPs trained on it.

Data Quality. The data quality and consistency were ensured through stringent design choices, including a series of filtering and sampling steps, as detailed in the Methods section. We conducted tests to ensure tight numerical convergence of the VASP settings used for the PBE-D3 calculations, as described in the Supplementary Information. Unrealistic frames were removed from relaxation trajectories by filtering based on energy, force, and stress. Consistent k-point grid density was maintained by volume filtering. Filtering based on molecular connectivity eliminated frames, in which molecular bonds were broken or unrealistically formed. The sampling strategy included frames from different stages of the relaxation trajectory, capturing configurations both far from and near equilibrium to achieve thorough sampling of the potential energy landscape. This sampling strategy provides a comprehensive representation of the system's evolution during the relaxation process, while avoiding unnecessary redundancy in the regions near equilibrium. Notably, our dataset exhibits high consistency between the initial and final frames of the sampled trajectories. Out of 230,168 structures, only 1,516 (0.7%) had the final frame with a different space group than that of the initial frame.

Chemical and Structural Diversity. To ensure broad applicability, the OMC25 dataset was designed to capture extensive chemical and structural diversity. As shown in Figure 2a, the OMC25 training split encompasses 12 elements most common in organic entries of the Cambridge Structural Database (CSD) [55]. Figure 2a also shows that the number of atoms in sampled molecules ranges from 4 to 164 with an average value of 42, and the number of atoms in the crystal unit cell from Genarris ranges from 12 to 300 with an average value of 130.

In terms of structural diversity, the OMC25 dataset includes 167 distinct space groups across all seven crystal systems. OMC25 features 22 space groups with more than 1% of structures versus only 10 in the CSD. In addition, there are notable differences in the prevalence of certain space groups in OMC25 compared to the CSD. Certain monoclinic space groups, such as Pc (No. 7) and P2 (No. 3), are significantly overrepresented in the OMC25 dataset, accounting for 11.2% and 4.9% of the starting molecular crystals, respectively. In contrast, these space groups are relatively rare in the CSD, with only 0.5% and 142 entries (out of over half a million), respectively. Similarly, the orthorhombic space group Pcc2 (No. 27) is represented at a frequency of 1.9% in the OMC25 dataset, compared to only 16 entries in the CSD. Furthermore, the tetragonal crystal system is significantly overrepresented in the OMC25 dataset, accounting for 16.5% of the structures, whereas it is relatively rare in the CSD, with only 1.5% representation. Conversely, the trigonal, hexagonal, and cubic systems are underrepresented in the OMC25 dataset, collectively accounting for only 0.7% of the structures, compared to 1.9% in the CSD. This difference may stem from not sampling Z=3 and from the lowered symmetry of the relaxed molecular structures in the OE62 dataset, which may have prevented structure generation in space groups with many special Wyckoff positions. The sampling strategy involved selecting the top six Z values from the CSD, which led to a diverse set of space groups and crystal systems represented in OMC25 (Figure 2b). The relationship between the number of molecules per unit cell and crystal symmetry is illustrated by the flow of sampled Z values into space groups and crystal systems in Figure 2c. This provides a visual illustration of the relationships between the number of molecules per unit cell and crystal symmetry.

Property Diversity. The analysis presented in Figure 3 reveals extensive sampling across large swaths of the potential energy landscape, which is crucial for training robust models. Figure 3a compares the density distributions of structures at various stages: initial Genarris generation, post-Rigid Press optimization, and final relaxation. The final relaxed structures closely match the density distribution of organic CSD entries, indicating realistic packing. As expected, the structures sampled after initial generation have the lowest density. Rigid Press effectively compresses the structures to achieve close-packing and significantly increases the density. The final relaxed structures, that closely follow CSD density distribution, are on average more dense than the as-generated structures but less dense than the structures compressed by Rigid Press. We sampled structures from both Genarris steps, ensuring that the final dataset includes almost equal number of both loosely and densely packed structures (Figure 3a). Starting from either loose or dense initial structures leads to different evolution of relaxation trajectories, depending on the interplay between the intensities of intramolecular and intermolecular forces. Figure 3b presents the distributions of the relaxation energy, maximum per-atom residual force, and maximum stress in the first and last frames of the relaxation trajectories that comprise the OMC25 dataset. The relaxation energy distribution highlights the extent of structural optimization, with an average difference of 0.7 eV/molecule between the initial and final structures. The force and stress

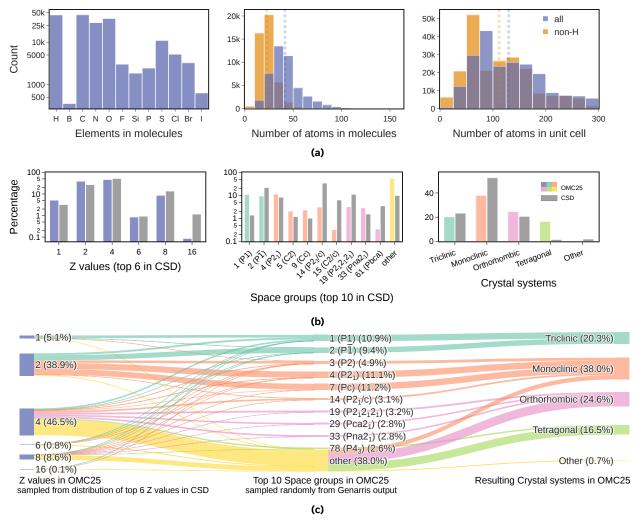


Figure 2 Description of the sampled molecules and molecular crystals before structure relaxations in the OMC25 training split: (a) elemental occurrences in molecules, distributions (histograms) and averages (vertical dotted lines) of the number of atoms with and without counting hydrogens in starting molecules and molecular crystals, (b) distributions of the top six most occurring Z values, the most occurring ten space groups, and four crystal systems in organic entries of CSD 6.00 [55] compared to OMC25, highlighted in yellow the collective distributions of the remaining space groups and crystal systems, (c) the Sankey diagram connecting Z values, the top ten most occurring space groups in OMC25, and resulting crystal systems, with the connecting links proportional to flow quantities. The coloring of space groups and connecting links is determined by crystal systems the space groups correspond to.

distributions demonstrate that the initial structures are far from equilibrium, while the final structures show tight convergence with the VASP settings chosen, indicating mostly finished successful relaxations. These trends are consistent across training, validation, and test subsets, confirming comprehensive sampling of the potential energy landscape and underscoring the reliability and robustness of the OMC25 dataset.

Model Performance. The primary goal of the OMC25 dataset is to enable accurate, transferable predictions of molecular crystal properties using machine learning interatomic potentials (MLIPs). To this end, we adopt a model-centric validation approach: the performance of state-of-the-art ML models trained on OMC25 serves as a direct, quantitative proxy for the dataset's informativeness, diversity, and completeness. High accuracy on relevant tasks (e.g., energy, force, and stress prediction) indicates that the dataset captures the essential physics and chemical diversity required for robust generalization. Conversely, poor model performance may reveal gaps, biases, or insufficient sampling. This approach complements traditional statistical or chemical diversity analyses and provides practical evidence of the dataset's utility for different applications and objectives such

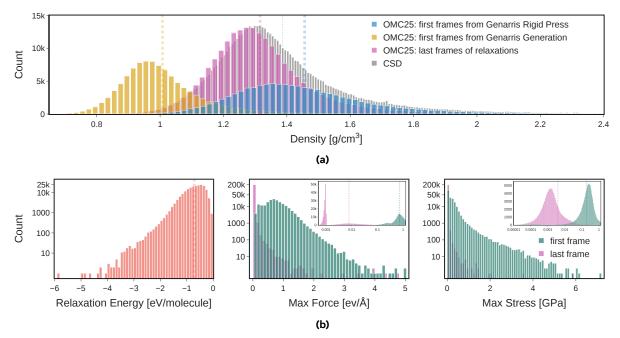


Figure 3 Comparison of properties from the first and last frames of relaxation trajectories of molecular crystals in the OMC25 training split: (a) showing density of the last frames and the first frames for different Genarris steps; overlayed is the distribution of densities of organic entries in CSD 6.00, (b) showing relaxation energy (the difference of energy between the initial and final frames), maximum per-atom residual force, and maximum unit cell stress. Average values are presented with vertical dotted lines.

as crystal structure prediction (CSP).

To technically validate the OMC25 dataset, we trained and benchmarked several state-of-the-art MLIPs, including UMA [20], eSEN [22], and EquiformerV2 [29]. These models represent the current frontier in molecular and materials modeling. Model and training parameters for eSEN and EquiformerV2 are presented in Table S2 and UMA model details can be found in [20]. All MLIPs are message-passing graph neural networks (GNNs) that operate on atomic graphs, with nodes representing atoms and edges representing neighboring atom pairs within a cutoff distance. We evaluate both energy-conserving models, which compute forces via automatic differentiation of the predicted energy (using PyTorch autograd [71]), and direct-force models, which predict forces as an explicit output. This distinction allows us to assess the impact of model architecture on predictive accuracy.

eSEN [22]: We selected the eSEN energy-conserving model due to its top ranking on the Matbench Discovery leaderboard [72] for inorganic materials and its strong performance for the Open Molecules 2025 (OMol25) dataset [31]. Given its success across both inorganic materials and molecular systems, we expect it to perform well for molecular crystals.

Universal Model for Atoms (UMA) [20]: UMA is a versatile model trained on diverse datasets and atomic system types, including the OMC25 dataset. Its architecture integrates the strengths of eSEN with a mixture of linear experts, enhancing adaptability and accuracy. We focus on the OMC task within UMA and evaluate energy-conserving UMA models of small and medium sizes to find an optimal balance between performance and complexity.

EquiformerV2 [29]: This direct-force GNN model incorporates transformer-inspired attention mechanisms and currently leads the Matbench Discovery leaderboard among direct-force models. We train EquiformerV2 using both a standard 6 Å cutoff and an extended 12 Å cutoff to compare the effects of model type (direct vs. energy-conserving) and receptive field size (i.e., the chemical environment an atom interacts with) on performance.

All MLIPs were evaluated on a held-out test subset, as well as on additional external benchmarks including the X23b benchmark [73] and the Schrödinger polymorph ranking task [74] described below. The results of

these evaluations are summarized in Table 3, with further details provided in the Supplementary Information.

X23b benchmark [73]: We benchmark our MLIPs using the X23b dataset, a revised version of the original X23 experimental set [75]. This dataset contains 23 small to medium-sized molecular crystals featuring a variety of intermolecular interactions such as van der Waals forces, hydrogen bonding, and mixed bonding types. The evaluation task involves predicting the unit cell volume and lattice energy of each crystal at 0 K, where lattice energy is defined as the cohesive energy of the crystal relative to the isolated (relaxed) molecule in the gas phase. Figure S2a shows the reference values alongside the predictions from our MLIPs for each system.

Schrödinger polymorph ranking [74]: To evaluate the MLIPs performance for ranking polymorphs, we tested the models on a recent polymorph dataset from [74]. As noted in [44], it is important to recognize that our MLIPs were trained on data generated with the PBE-D3 functional [52, 53], whereas the reference polymorph energies were computed using the r²SCAN-D3 meta-GGA functional [53, 76], which is considered more accurate. This discrepancy introduces some inherent limitations in the benchmark. For each system, we computed energy and rank correlation metrics, which were then averaged to produce the final evaluation scores. Distributions of these relative lattice energy metrics are presented in Figure S2b.

Our results show that the OMC25 dataset can be used to train highly accurate machine learning models, with low energy, force, and stress MAEs, for the X23b evaluation. The trained MLIPs are also well-suited for the polymorph ranking task, even considering that the reference dataset was built at a higher level of density functional theory. It is important to note that the direct force models tend to perform poorly for relaxation tasks, and thus are not recommended for such applications.

To emphasize the critical role of crystal-specific data, we conducted a comparative evaluation of identical MLIP architectures trained on the OMC25 crystal dataset and the OMol25 molecular dataset [31]. We note that the OMol25 dataset was acquired using the ω B97M-V range-separated hybrid meta-GGA functional with the VV10 non-local treatment of dispersion interactions [77–79], which is significantly more accurate than PBE-D3. In addition, the OMol25 dataset was acquired using Gaussian basis sets without pseudopotentials of light atoms, although this is expected to have a more minor effect than the accuracy of the exchange-correlation functional and dispersion method. OMol25 is intended to be a general purpose molecular dataset and therefore lacks data on large clusters of organic molecules at various separations (though it does contain solvated systems and protein-ligand pockets) as those in OMC25. The differences in the performance of models trained on OMC25 vs. OMol25 may thus be attributed to the different DFT settings, the use of periodic vs. non-periodic codes, and the distributional shift of the underlying data. Our results reveal that model performance varies

Table 3 MLIP evaluations: validation and test metrics, as well as X23b and Schrödinger polymorph ranking evaluations. The energy, force, and stress errors are presented as mean absolute errors (MAEs) for the validation and test metrics. For the X23b evaluation, the energy represents the lattice energy of the crystal, and the mean absolute percentage error (MAPE) is reported for the molar volume. For the Schrödinger polymorph ranking evaluation, the energies are normalized to the number of molecules in the unit cell. The bolded values show the best performing models, where the evaluation followed the community standards.

	so		I	/alidatio	n		Test		X231	b [73]	Sc	chröding	er
	eter	_									polymo	rph rank	ing [74]
Model	Number of Parameters	Conserving model	Energy \(\int \] [meV/atom]	Forces \(\) [meV/\hat{A}]	Stress \downarrow [meV/ 3]	Energy \(\) [meV/atom]	Forces \downarrow [meV/Å]	Stress \downarrow [meV/ 3]	Lattice Energy MAE [kcal/mol] \undergale	Volume $MAPE[\%] \downarrow$	Rel. Energy MAE [kcal/mol] \upspace	Correlation Pearson ↑	Rank corelation Spearman ↑
UMA-S-1.1 (OMC) [20]	6M [†]	1	1.05	5.18	0.95	1.03	5.04	0.93	2.21	6.01	0.35	0.80	0.74
UMA-M-1.1 (OMC) [20]	50M [†]	1	0.86	2.92	0.92	0.84	2.83	0.90	1.94	5.78	0.44	0.73	0.68
eSEN-S-OMC [22]	6M	1	1.06	5.58	0.96	1.05	5.39	0.94	3.38	5.58	1.04	0.76	0.72
eqV2-S-OMC (6 Å) [‡] [29]	31M	X	0.61	3.89	0.11	0.70	3.87	0.11	8.87	4.02	0.39	0.78	0.74
eqV2-S-OMC (12 Å) [‡] [29	31M	Х	0.62	3.79	0.10	0.67	3.78	0.10	9.09	2.50	0.34	0.79	0.74

[†] Reported is the number of active parameters during inference, which is lower than the total number of parameters used to train UMA models [20].

† For the X23b benchmark, the single point energies of starting molecular structures in the gas phase were taken as the reference for lattice energy calculations. For the Schrödinger polymorph ranking, the single point energies of starting crystals were taken as the reference for relative energy

depending on the evaluation task (Table S1). For the X23b lattice energies, most UMA models achieve comparable accuracy, while the X23b unit cell volumes are best predicted with the OMol task of UMA. For the Schrödinger polymorph ranking task, UMA models with the OMC task show superior performance in energy metrics. This highlights that molecular and crystal datasets offer complementary advantages. This also underscores the necessity of including crystal data to fully capture the complexities relevant to molecular crystal modeling. Additional details are provided in the Supplementary Information.

Limitations and Future Directions. While the OMC25 dataset and the associated MLIPs offer significant advances, several limitations remain that warrant discussion and guide future research.

First, we limited OMC25 to single-component pristine organic molecular crystals with one molecule in the asymmetric unit (Z'=1). However, other classes of molecular crystals—such as co-crystals, multi-component systems, hydrates, solvates, metal-organic frameworks, and disordered systems—are of great practical interest and involve unique intermolecular interactions that remain to be explored. Future work should also prioritize richer elemental diversity beyond that in the OE62 dataset [51]. Additionally, this study used a starting single geometry-optimized molecular conformer from the OE62 dataset (except for a handful of cases). Although DFT relaxes the atomic positions in a crystal, there is no guarantee that, in practice, the resulting structures will contain highly different molecular conformers. Incorporating multiple conformers in crystal generation will be an important step toward better modeling conformer interactions.

Second, the level of density functional theory (PBE-D3 [52, 53]) employed for structure relaxations was deemed sufficient; nonetheless, several studies indicate that higher-level approaches—such as hybrid functionals like PBE0 [80] and meta-GGA functionals like r²SCAN [81]—along with more advanced dispersion corrections (e.g., VV10 [77], XDM [82], TS [59], and MBD [83, 84])—are essential for accurately capturing crystal energetics [73, 75, 81, 85–89]. These enhancements are particularly important for CSP tasks, where high precision is required to distinguish polymorphs that differ by only a few kJ/mol [90–95].

Third, a known limitation of the current MLIPs lies in capturing long-range interactions. Message-passing layers can extend the effective receptive field beyond the cutoff distance, depending on the maximum number of neighbors allowed in the graph neural network (GNN). However, challenges persist in accurately modeling interactions in systems, such as organic-organic interfaces, with very long-ranged interactions, extending far beyond the cutoff radius. Additionally, the test and validation sets used in this study closely resemble the training data, limiting the ability to fully assess the generalizability of the MLIPs. To more rigorously evaluate these models, future benchmarks should incorporate both in-distribution and out-of-distribution test sets [96] to better measure their transferability and robustness.

In summary, the OMC25 dataset is a unique resource, providing high-quality property labels for a rich and diverse set of molecular crystal structures. The MLIPs trained on OMC25 demonstrate impressive accuracy in predicting molecular crystal structures and relative energies [97]. Together, they pave the way for accelerated simulations and more reliable molecular crystal structure and property predictions, opening new frontiers in molecular crystal research.

Code Availability

The random molecular crystal structure generation was performed using Genarris 3.0 package [54], which is available at https://github.com/Yi5817/Genarris and https://www.noamarom.com/software/genarris under the BSD 3-Clause license. All density functional theory (DFT) calculations were carried out with the Vienna Ab initio Simulation Package (VASP) [62–64]. The scripts to generate VASP input files with OMC25 settings are publicly available at https://github.com/facebookresearch/fairchem/tree/main/src/fairchem/data/omc. The machine learning interatomic potentials trained on the OMC25 dataset are accessible from HuggingFace: eSEN model [22] at https://huggingface.co/facebook/OMC25 and UMA models [20] at https://huggingface.co/facebook/UMA. We provide the pretrained models with a commercially permissive license (with some geographic and acceptable use restrictions). The code containing all necessary information for reproducing our training and evaluation results is publicly available at https://github.com/facebookresearch/fairchem. We encourage users to cite this paper when using the dataset or pretrained models for molecular crystals in their research.

Competing Interests

The authors declare no competing interests.

Acknowledgments

N.M. acknowledges support from the National Science Foundation (NSF) Designing Materials to Revolutionize and Engineer our Future (DMREF) program via award DMR-2323749. Y.Y. acknowledges support from the Frontera Computational Science Fellowship awarded by the Texas Advanced Computing Center (TACC).

Author Contributions

V.G. generated the dataset; L.B.L. conducted the DFT convergence study; Y.Y. helped with configuring Genarris 3.0; everyone was involved in training, evaluating, and/or improving models; M.U. supervised the project in the initial stages; A.B., C.L.Z., N.M., Z.W.U., and A.S. supervised the entire project; V.G. prepared and everyone contributed to the review of the manuscript.

References

- [1] Brittain, H. G. Polymorphism in Pharmaceutical Solids (Marcel Dekker, New York, 1999).
- [2] Vishweshwar, P., McMahon, J. A., Bis, J. A., and Zaworotko, M. J. Pharmaceutical co-crystals. *Journal of Pharmaceutical Sciences* 95, 499-516 (2006).
- [3] Morissette, S. L., Almarsson, Ö., Peterson, M. L., Remenar, J. F., Read, M. J., Lemmo, A. V., Ellis, S., Cima, M. J., and Gardner, C. R. High-throughput crystallization: polymorphs, salts, co-crystals and solvates of pharmaceutical solids. Advanced Drug Delivery Reviews 56, 275–300 (2004).
- [4] Rodríguez-Hornedo, N., Nehm, S. J., and Jayasankar, A. Cocrystals: design, properties and formation mechanisms. *Encyclopedia of Pharmaceutical Technology* 1, 615–635 (2007).
- [5] Blagden, N., de Matas, M., Gavan, P. T., and York, P. Crystal engineering of active pharmaceutical ingredients to improve solubility and dissolution rates. *Advanced Drug Delivery Reviews* **59**, 617–630 (2007).
- [6] Wang, C., Dong, H., Jiang, L., and Hu, W. Organic semiconductor crystals. Chemical Society Reviews 47, 422–500 (2018).
- [7] Dong, H., Fu, X., Liu, J., Wang, Z., and Hu, W. 25th anniversary article: key points for high-mobility organic field-effect transistors. *Advanced Materials* **25**, 6158–6183 (2013).
- [8] Wang, Y., Sun, L., Wang, C., Yang, F., Ren, X., Zhang, X., Dong, H., and Hu, W. Organic crystalline materials in flexible electronics. *Chemical Society Reviews* 48, 1492–1530 (2019).
- [9] Zhang, X., Dong, H., and Hu, W. Organic semiconductor single crystals for electronics and photonics. Advanced Materials 30, 1801048 (2018).
- [10] Hao, Z., and Iqbal, A. Some aspects of organic pigments. Chemical Society Reviews 26, 203–213 (1997).
- [11] Yu, Q., Aguila, B., Gao, J., Xu, P., Chen, Q., Yan, J., Xing, D., Chen, Y., Cheng, P., Zhang, Z., and Ma, S. Photomechanical organic crystals as smart materials for advanced applications. *Chemistry–A European Journal* **25**, 5611–5622 (2019).
- [12] Fang, H.-H., Yang, J., Feng, J., Yamao, T., Hotta, S., and Sun, H.-B. Functional organic single crystals for solid-state laser applications. *Laser & Photonics Reviews* **8**, 687–715 (2014).
- [13] Tozawa, T., Jones, J. T., Swamy, S. I., Jiang, S., Adams, D. J., Shakespeare, S., Clowes, R., Bradshaw, D., Hasell, T., Chong, S. Y., Tang, C., Thompson, S., Parker, J., Trewin, A., Bacsa, J., Slawin, A., Alexandra M Z amd Steiner, and Cooper, A. I. Porous organic cages. *Nature Materials* 8, 973–978 (2009).
- [14] Hunger, K., and Schmidt, M. U. Industrial organic pigments: production, crystal structures, properties, applications (John Wiley & Sons, 2019).

- [15] Niu, X., Yang, R., Zhang, H., and Yang, J. Crystal engineering in the development of improved pesticide products. Advanced Agrochem 1, 39–60 (2022).
- [16] Lee, A. Y., Erdemir, D., and Myerson, A. S. Crystal polymorphism in chemical process development. Annual Review of Chemical and Biomolecular Engineering 2, 259–280 (2011).
- [17] Bernstein, J. Polymorphism in Molecular Crystals, vol. 30 (Oxford University Press, 2020).
- [18] Wines, D., and Choudhary, K. CHIPS-FF: Evaluating universal machine learning force fields for material properties. ACS Materials Letters 7, 2105–2114 (2025).
- [19] Loew, A., Sun, D., Wang, H.-C., Botti, S., and Marques, M. A. Universal machine learning interatomic potentials are ready for phonons. *npj Computational Materials* **11**, 1–8 (2025).
- [20] Wood, B. M., Dzamba, M., Fu, X., Gao, M., Shuaibi, M., Barroso-Luque, L., Abdelmaqsoud, K., Gharakhanyan, V., Kitchin, J. R., Levine, D. S., Michel, K., Sriram, A., Cohen, T., Das, A., Rizvi, A., Sahoo, S. J., Ulissi, Z. W., and Zitnick, C. L. UMA: A Family of Universal Models for Atoms. arXiv preprint arXiv:2506.23971 (2025).
- [21] Batatia, I., Kovacs, D. P., Simm, G., Ortner, C., and Csányi, G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, vol. 35, 11423–11436 (2022).
- [22] Fu, X., Wood, B. M., Barroso-Luque, L., Levine, D. S., Gao, M., Dzamba, M., and Zitnick, C. L. Learning smooth and expressive interatomic potentials for physical property prediction. In *International Conference on Machine Learning* (2025).
- [23] Schütt, K., Kindermans, P.-J., Sauceda Felix, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, vol. 30 (2017).
- [24] Sriram, A., Das, A., Wood, B. M., and Zitnick, C. L. Towards training billion parameter graph neural networks for atomic simulations. In *International Conference on Learning Representations* (2022).
- [25] Gasteiger, J., Becker, F., and Günnemann, S. GemNet: Universal directional graph neural networks for molecules. In Advances in Neural Information Processing Systems, vol. 34, 6790–6802 (2021).
- [26] Gasteiger, J., Shuaibi, M., Sriram, A., Günnemann, S., Ulissi, Z. W., Zitnick, C. L., and Das, A. GemNet-OC: Developing graph neural networks for large and diverse molecular simulation datasets. *Transactions on Machine Learning Research (TMLR)* (2022).
- [27] Schütt, K., Unke, O., and Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning*, 9377–9388 (PMLR, 2021).
- [28] Passaro, S., and Zitnick, C. L. Reducing SO(3) convolutions to SO(2) for efficient equivariant GNNs. In International Conference on Machine Learning, 27420–27438 (PMLR, 2023).
- [29] Liao, Y.-L., Wood, B. M., Das, A., and Smidt, T. EquiformerV2: Improved equivariant transformer for scaling to higher-degree representations. In *International Conference on Learning Representations* (2024).
- [30] Jacobs, R. et al. A practical guide to machine learning interatomic potentials-status and future. Current Opinion in Solid State and Materials Science 35, 101214 (2025).
- [31] Levine, D. S., Shuaibi, M., Spotte-Smith, E. W. C., Taylor, M. G., Hasyim, M. R., Michel, K., Batatia, I., Csányi, G., Dzamba, M., Eastman, P., Frey, N. C., Fu, X., Gharakhanyan, V., Krishnapriyan, A. S., Rackers, J. A., Raja, S., Rizvi, A., Rosen, A. S., Ulissi, Z., Vargas, S., Zitnick, C. L., Blau, S. M., and Wood, B. M. The Open Molecules 2025 (OMol25) dataset, evaluations, and models. arXiv preprint arXiv:2505.08762 (2025).
- [32] Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* 1, 1–7 (2014).
- [33] Blum, L. C., and Reymond, J.-L. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society* **131**, 8732–8733 (2009).
- [34] Axelrod, S., and Gomez-Bombarelli, R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. Scientific Data 9, 185 (2022).
- [35] Eastman, P., Behara, P. K., Dotson, D. L., Galvelis, R., Herr, J. E., Horton, J. T., Mao, Y., Chodera, J. D., Pritchard, B. P., Wang, Y., De Fabritiis, G., and Markland, T. E. SPICE, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data* 10, 11 (2023).

- [36] Schreiner, M., Bhowmik, A., Vegge, T., Busk, J., and Winther, O. Transition1x-a dataset for building generalizable reactive machine learning potentials. *Scientific Data* 9, 779 (2022).
- [37] Smith, J. S., Zubatyuk, R., Nebgen, B., Lubbers, N., Barros, K., Roitberg, A. E., Isayev, O., and Tretiak, S. The ANI-1cx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Scientific Data* 7, 134 (2020).
- [38] Barroso-Luque, L., Shuaibi, M., Fu, X., Wood, B. M., Dzamba, M., Gao, M., Rizvi, A., Zitnick, C. L., and Ulissi, Z. W. Open Materials 2024 (OMat24) inorganic materials dataset and models. arXiv preprint arXiv:2410.12771 (2024).
- [39] Chanussot, L., Das, A., Goyal, S., Lavril, T., Shuaibi, M., Riviere, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., Palizhati, A., Sriram, A., Wood, B., Yoon, J., Parikh, C. L., Devi amd Zitnick, and Ulissi, Z. Open Catalyst 2020 (OC20) dataset and community challenges. ACS Catalysis 11, 6059-6072 (2021).
- [40] Tran, R., Lan, J., Shuaibi, M., Wood, B. M., Goyal, S., Das, A., Heras-Domingo, J., Kolluru, A., Rizvi, A., Shoghi, N., Sriram, A., Therrien, F., Abed, J., Voznyy, O., Sargent, E. H., Ulissi, Z., and Zitnick, C. L. The Open Catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts. ACS Catalysis 13, 3066–3084 (2023).
- [41] Sriram, A., Choi, S., Yu, X., Brabson, L. M., Das, A., Ulissi, Z., Uyttendaele, M., Medford, A. J., and Sholl, D. S. The Open DAC 2023 dataset and challenges for sorbent discovery in direct air capture. ACS Central Science 10, 923–941 (2024).
- [42] Kaplan, A. D., Liu, R., Qi, J., Ko, T. W., Deng, B., Riebesell, J., Ceder, G., Persson, K. A., and Ong, S. P. A foundational potential energy surface dataset for materials. arXiv preprint arXiv:2503.04070 (2025).
- [43] Žugec, I., Geilhufe, R. M., and Lončarić, I. Global machine learning potentials for molecular crystals. The Journal of Chemical Physics 160, 154106 (2024).
- [44] Mann, E. L., Wagen, C. C., Vandezande, J. E., Wagen, A. M., and Schneider, S. C. Egret-1: Pretrained neural network potentials for efficient and accurate bioorganic simulation. arXiv preprint arXiv:2504.20955 (2025).
- [45] Kovács, D. P., Moore, J. H., Browning, N. J., Batatia, I., Horton, J. T., Pu, Y., Kapil, V., Witt, W. C., Magdau, I.-B., Cole, D. J., and Csányi, G. MACE-OFF: Short-range transferable machine learning force fields for organic molecules. *Journal of the American Chemical Society* 147, 17598–17611 (2025).
- [46] Smith, J. S., Isayev, O., and Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical Science* 8, 3192–3203 (2017).
- [47] Anstine, D. M., Zubatyuk, R., and Isayev, O. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chemical Science* **16**, 10228–10244 (2025).
- [48] Taylor, C. R., Butler, P. W., and Day, G. M. Predictive crystallography at scale: mapping, validating, and learning from 1000 crystal energy landscapes. *Faraday Discussions* **256**, 434–458 (2025).
- [49] Borysov, S. S., Geilhufe, R. M., and Balatsky, A. V. Organic materials database: An open-access online database for data mining. *PLOS One* **12**, e0171501 (2017).
- [50] Weber, J. L., Guha, R. D., Agarwal, G., Wei, Y., Fike, A. A., Xie, X., Stevenson, J., Leswing, K., Halls, M. D., Abel, R., and Jacobson, L. D. Efficient long-range machine learning force fields for liquid and materials properties. arXiv preprint arXiv:2505.06462 (2025).
- [51] Stuke, A., Kunkel, C., Golze, D., Todorović, M., Margraf, J. T., Reuter, K., Rinke, P., and Oberhofer, H. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Scientific Data* 7, 58 (2020).
- [52] Perdew, J. P., Burke, K., and Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters* 77, 3865 (1996).
- [53] Grimme, S., Antony, J., Ehrlich, S., and Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. The Journal of Chemical Physics 132, 154104 (2010).
- [54] Yang, Y., Tom, R., Wui, J. A., Moussa, J. E., and Marom, N. Genarris 3.0: Generating close-packed molecular crystal structures with rigid press. ChemRxiv preprint chemrxiv:2025-046zn (2025).
- [55] Groom, C. R., Bruno, I. J., Lightfoot, M. P., and Ward, S. C. The Cambridge structural database. Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials 72, 171–179 (2016).

- [56] Blum, V., Gehrke, R., Hanke, F., Havu, P., Havu, V., Ren, X., Reuter, K., and Scheffler, M. Ab initio molecular simulations with numeric atom-centered orbitals. Computer Physics Communications 180, 2175–2196 (2009).
- [57] Ren, X., Rinke, P., Blum, V., Wieferink, J., Tkatchenko, A., Sanfilippo, A., Reuter, K., and Scheffler, M. Resolution-of-identity approach to Hartree–Fock, hybrid density functionals, RPA, MP2 and GW with numeric atom-centered orbital basis functions. New Journal of Physics 14, 053020 (2012).
- [58] Zhang, I. Y., Ren, X., Rinke, P., Blum, V., and Scheffler, M. Numeric atom-centered-orbital basis sets with valence-correlation consistency from H to Ar. New Journal of Physics 15, 123033 (2013).
- [59] Tkatchenko, A., and Scheffler, M. Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. *Physical Review Letters* 102, 073005 (2009).
- [60] Tom, R., Rose, T., Bier, I., O'Brien, H., Vázquez-Mayagoitia, Á., and Marom, N. Genarris 2.0: A random structure generator for molecular crystals. Computer Physics Communications 250, 107170 (2020).
- [61] Bier, I., and Marom, N. Machine learned model for solid form volume estimation based on packing-accessible surface and molecular topological fragments. The Journal of Physical Chemistry A 124, 10330–10345 (2020).
- [62] Kresse, G., and Hafner, J. Ab initio molecular dynamics for liquid metals. Physical Review B 47, 558 (1993).
- [63] Kresse, G., and Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B* 54, 11169 (1996).
- [64] Kresse, G., and Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. Physical Review B 59, 1758 (1999).
- [65] Blöchl, P. E. Projector augmented-wave method. Physical Review B 50, 17953 (1994).
- [66] Ranganathan, S. I., and Ostoja-Starzewski, M. Universal elastic anisotropy index. Physical Review Letters 101, 055504 (2008).
- [67] Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A., and Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* 68, 314–319 (2013).
- [68] Hagberg, A. A., Schult, D. A., and Swart, P. J. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the Python in Science Conference*, 11–15 (SciPy, 2008).
- [69] Chu, H. Lightning Memory-Mapped Database (LMDB) (2009). Part of the OpenLDAP project.
- [70] Larsen, A. H. et al. The atomic simulation environment—a Python library for working with atoms. Journal of Physics: Condensed Matter 29, 273002 (2017).
- [71] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. In *Conference on Neural Information Processing Systems* (2017).
- [72] Riebesell, J., Goodall, R. E., Benner, P., Chiang, Y., Deng, B., Ceder, G., Asta, M., Lee, A. A., Jain, A., and Persson, K. A. A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence* 7, 836–847 (2025).
- [73] Dolgonos, G. A., Hoja, J., and Boese, A. D. Revised values for the X23 benchmark set of molecular crystals. Physical Chemistry Chemical Physics 21, 24333–24344 (2019).
- [74] Zhou, D., Bier, I., Santra, B., Jacobson, L. D., Wu, C., Garaizar Suarez, A., Almaguer, B. R., Yu, H., Abel, R., Friesner, R. A., and Wang, L. A robust crystal structure prediction method to support small molecule drug development with large scale validation and blind study. *Nature Communications* 16, 2210 (2025).
- [75] Reilly, A. M., and Tkatchenko, A. Understanding the role of vibrations, exact exchange, and many-body van der Waals interactions in the cohesive properties of molecular crystals. The Journal of Chemical Physics 139, 024705 (2013).
- [76] Furness, J. W., Kaplan, A. D., Ning, J., Perdew, J. P., and Sun, J. Accurate and numerically efficient r2SCAN meta-generalized gradient approximation. The Journal of Physical Chemistry Letters 11, 8208–8215 (2020).
- [77] Mardirossian, N., and Head-Gordon, M. ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. The Journal of Chemical Physics 144, 214110 (2016).

- [78] Hellweg, A., and Rappoport, D. Development of new auxiliary basis functions of the Karlsruhe segmented contracted basis sets including diffuse basis functions (def2-SVPD, def2-TZVPPD, and def2-QVPPD) for RI-MP2 and RI-CC calculations. *Physical Chemistry Chemical Physics* 17, 1010–1017 (2015).
- [79] Rappoport, D., and Furche, F. Property-optimized gaussian basis sets for molecular response calculations. The Journal of Chemical Physics 133, 134105 (2010).
- [80] Adamo, C., and Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of Chemical Physics* **110**, 6158–6170 (1999).
- [81] Grimme, S., Hansen, A., Ehlert, S., and Mewes, J.-M. r2SCAN-3c: A "swiss army knife" composite electronic-structure method. The Journal of Chemical Physics 154, 064103 (2021).
- [82] Becke, A. D., and Johnson, E. R. Exchange-hole dipole moment and the dispersion interaction. The Journal of Chemical Physics 122, 154104 (2005).
- [83] Tkatchenko, A., DiStasio Jr, R. A., Car, R., and Scheffler, M. Accurate and efficient method for many-body van der Waals interactions. *Physical Review Letters* 108, 236402 (2012).
- [84] Ambrosetti, A., Reilly, A. M., DiStasio, R. A., and Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. The Journal of Chemical Physics 140, 18A508 (2014).
- [85] Otero-De-La-Roza, A., and Johnson, E. R. A benchmark for non-covalent interactions in solids. The Journal of Chemical Physics 137, 054103 (2012).
- [86] Marom, N., DiStasio Jr, R. A., Atalla, V., Levchenko, S., Reilly, A. M., Chelikowsky, J. R., Leiserowitz, L., and Tkatchenko, A. Many-body dispersion interactions in molecular crystal polymorphism. *Angewandte Chemie International Edition* 52, 6629–6632 (2013).
- [87] Beran, G. J. Modeling polymorphic molecular crystals with electronic structure theory. Chemical Reviews 116, 5567–5613 (2016).
- [88] Hermann, J., DiStasio Jr, R. A., and Tkatchenko, A. First-principles models for van der Waals interactions in molecules and materials: Concepts, theory, and applications. *Chemical Reviews* 117, 4714–4758 (2017).
- [89] O'Connor, D., Bier, I., Hsieh, Y.-T., and Marom, N. Performance of dispersion-inclusive density functional theory methods for energetic materials. *Journal of Chemical Theory and Computation* **18**, 4456–4471 (2022).
- [90] Whittleton, S. R., Otero-De-La-Roza, A., and Johnson, E. R. Exchange-hole dipole dispersion model for accurate energy ranking in molecular crystal structure prediction. *Journal of Chemical Theory and Computation* 13, 441–450 (2017).
- [91] Whittleton, S. R., Otero-De-La-Roza, A., and Johnson, E. R. Exchange-hole dipole dispersion model for accurate energy ranking in molecular crystal structure prediction ii: Nonplanar molecules. *Journal of Chemical Theory* and Computation 13, 5332–5342 (2017).
- [92] Hoja, J., and Tkatchenko, A. First-principles stability ranking of molecular crystal polymorphs with the DFT+ MBD approach. Faraday Discussions 211, 253–274 (2018).
- [93] Hoja, J., Ko, H.-Y., Neumann, M. A., Car, R., DiStasio Jr, R. A., and Tkatchenko, A. Reliable and practical computational description of molecular crystal polymorphs. *Science Advances* **5**, eaau3338 (2019).
- [94] Price, A. J., Mayo, R. A., Otero-de-la Roza, A., and Johnson, E. R. Accurate and efficient polymorph energy ranking with XDM-corrected hybrid DFT. *CrystEngComm* **25**, 953–960 (2023).
- [95] Beran, G. J. Frontiers of molecular crystal structure prediction for pharmaceuticals and functional organic materials. Chemical Science 14, 13290–13312 (2023).
- [96] Omee, S. S., Fu, N., Dong, R., Hu, M., and Hu, J. Structure-based out-of-distribution (OOD) materials property prediction: a benchmark study. npj Computational Materials 10, 144 (2024).
- [97] Gharakhanyan, V., Yang, Y., Barroso-Luque, L., Shuaibi, M., Levine, D. S., Michel, K., Bernat, V., Dzamba, M., Fu, X., Gao, M., Liu, X., Noori, K., Purvis, L. J., Rao, T., Wood, B. M., Rizvi, A., Uyttendaele, M., Ouderkirk, A. J., Daraio, C., Zitnick, C. L., Boromand, A., Marom, N., Ulissi, Z. W., and Sriram, A. FastCSP: Accelerated Molecular Crystal Structure Prediction with Universal Model for Atoms. arXiv preprint (2025).
- [98] Gupta, N., Narayanan, I., Handa, S., Chakraborti, S., Thapar, P., Shan, B., Rao, A., Liu, Y., Wang, P., Wu, Y., Gao, Q., Cheng, C. C.-C., You, S., Huang, L., Fan, J., Yu, K., Lin, K., Mu, T., Malani, P., Wang, H., Lu, T.,

- and Zhang, P. Dynamic idle resource leasing to safely oversubscribe capacity at Meta. In *Proceedings of the 2024 ACM Symposium on Cloud Computing*, SoCC'24, 792–810 (2024).
- [99] Ganose, A. M. et al. Atomate2: Modular workflows for materials science. Digital Discovery 4, 1944–1973 (2025).
- [100] Hewat, A., and Riekel, C. Crystal structure of deuteroammonia between 2 and 180 K by neutron powder profile refinement. Acta Crystallographica Section A: Foundations and Advances (Denmark) 35, 569–571 (1979).
- [101] Simon, A., and Peters, K. Single-crystal refinement of the structure of carbon dioxide. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **36**, 2750–2751 (1980).
- [102] Hammer, B., Hansen, L. B., and Nørskov, J. K. Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals. *Physical Review B* **59**, 7413 (1999).

Supplementary Information

Genarris Details

```
[master]
name = <svstem CSD reference code>
molecule_path = <path to molecular geometry file>
Z = <sampled Z number>
log_level = debug
restart = True
[generation]
num\_structures\_per\_spg = 2
specific_radius_proportion = 0.95
natural_cutoff_mult = 1.5
tol = 0.01
spg_distribution_type = standard
max_attempts_per_spg = 1000000
unit_cell_volume_mean = predict
volume_mult = 1.25
max_attempts_per_volume = 100000
generation_type = crystal
[symm_rigid_press]
sr = 0.85
natural\_cutoff\_mult = 1.2
int scale = 0.1
method = BFGS
tol = 1e-3
maxiter = 5000
debug_flag = True
```

Listing S1 The base Genarris 3.0 [54] configuration file used for random molecular crystal generation.

DFT Details

VASP version 6.3 [62–64] was used for all calculations. Calculations were executed across various machine sizes and processor types using Elastic Compute within Meta's private cloud [98], with parallelization parameters (such as NCORE and the number of MPI ranks) adjusted to the architecture of each server. Since these servers running on Elastic Compute can be preempted at any time, a single VASP calculation could be stopped and restarted many times before convergence criteria had been satisfied. With each restart, the POSCAR file was replaced with the CONTCAR file that was present when VASP stopped last and, as WAVECAR files were not written during these calculations, wavefunctions were re-initialized each time VASP was started. VASP inputs were generated using RelaxSetGenerator class from atomate [99]. For all structures, VASP 5.4 PBE pseudopotentials were selected, as they are suitable for the non-exotic elements present in this dataset. The atomic positions and lattice vectors were relaxed until the maximum per-atom residual forces fell below 0.001 eV/Å, or the relaxation process exceeded 1.500 steps for most crystals, although a small, randomly selected subset (around 17% of structures) was allowed to relax up to 3,000 ionic steps. The total energy convergence tolerance was set to 10^{-6} eV, and the plane-wave energy cut-off was fixed at 520 eV, based on the recommended ENCUT=1.3×ENMAX with maximum ENMAX=400 for the elements in our dataset. A maximum of 200 electronic self-consistency steps were allowed. The default k-point density in PYMATGEN (using the Γ -centered strategy) was applied. Relaxation outputs were parsed with ASE [70] and validated using several simple consistency checks.

```
ADDGRID = True
ALGO = Normal
EDIFF = 1e-06
EDIFFG = -0.001
ENAUG = 1360
ENCUT = 520
GGA = Pe
IBRION = 2
ISIF = 3
ISMEAR = 0
ISPIN = 1
IVDW = 11
LASPH = True
LMIXTAU = True
LORBIT = 11
LREAL = False
NELM = 200
NELMDL = -10
NSW = 1500
PREC = Normal
SIGMA = 0.1
```

Listing S2 Example INCAR settings for the DFT relaxations with VASP.

DFT Convergence

A convergence study was conducted on 500 representative molecular crystals to validate the VASP convergence criteria for total energy. Due to computational limitations, only 484 and 73 structures were used for energy tolerance and k-point density calculations, respectively. The results are presented in Figure S1. The value used for energy tolerance (10^{-6} eV) was found to be sufficiently converged for the structural relaxations reported in this work. The final energy values showed slower convergence with respect to k-point density. The error statistics for k-point meshes generated by PYMATGEN was deemed sufficient for the calculations performed. This study validates that the chosen thresholds are sufficient for the calculations.

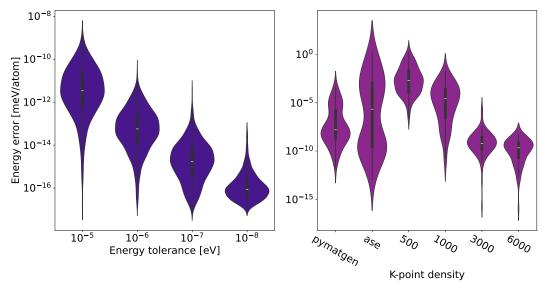


Figure S1 Convergence study for the DFT settings used in this work. We show the effect of each parameter on the energy of the structure compared to the most tightly converged settings set to 10^{-9} eV energy tolerance and 9,000 k-point density per atom.

Evaluation Details

X23b benchmark [73]: All starting molecular crystal structures were obtained from [73] as low temperature polymorphs and that of ammonia and carbon dioxide crystals were taken from [100] and [101], respectively. We used lattice energy reference values from [73] but the volumes were taken from [81] to remove the effects of the revised Perdew, Burke, and Ernzerhof (RPBE) [102] level of theory calculations on volumes leading to unrealistically large values for selected systems. We used ASE [70] and relaxed structures until the maximum per-atom residual forces were smaller than 0.001 eV/Å, or for a maximum of 5,000 steps, constraining the relaxation to the experimental space groups. For the direct-force models, we took the single point energies of starting molecular structures in the gas phase as the reference for lattice energy calculations.

Schrödinger polymorph ranking [74]: For each set of polymorphs of 66 systems studied, we used ASE [70] and relaxed structures with the energy-conserving models until the maximum per-atom residual forces were smaller than 0.01 eV/Å, or for a maximum of 5,000 steps. For the direct-force models, we took the single point energies. Energy and rank correlation metrics were first computed for each system (where appropriate) and then averaged to derive the final evaluation metrics. We note that OMC25 training split contains putative structures of 24 out of 66 systems included in this benchmark.

Table S1 Extended MLIP evaluation results for OMC25 and OMol25 [31] models: validation and test metrics, as well as X23b benchmark and Schrödinger polymorph ranking task. For UMA models [20], we used both the OMC and OMol tasks. The bolded values show the best performing models.

				Validation			Test		X23b [73]		Schrödinger				
	eter										polymorph ranking [74]				
Model	Number of Parameters	Conserving model	Energy ↓ [meV/atom]	Forces \downarrow [meV/Å]	Stress \downarrow $[meV/\hbar^3]$	Energy ↓ [meV/atom]	Forces \downarrow [meV/ \mathring{A}]	Stress \downarrow [meV/ 3]	Lattice Energy MAE [kcal/mol] ↓	Volume $MAPE [\%] \downarrow$	Rel. Energy MAE [kcal/mol] \(\dagger	RMSE [kcal/mol] 🕹	Correlation Pearson ↑	Rank corelation Kendall ↑	Rank corelation Spearman ↑
UMA-S-1.1 (OMC) [20]	6M [†]	1	1.05	5.18	0.95	1.03	5.04	0.93	2.21	6.01	0.35	0.43	0.80	0.60	0.74
UMA-S-1.1 (OMol) [20]	$6M^{\dagger}$	1	-	-	-	-	-	-	2.21	2.23	0.55	0.68	0.76	0.59	0.74
UMA-M-1.1 (OMC) [20]	50M [†]	1	0.86	2.92	0.92	0.84	2.83	0.90	1.94	5.78	0.44	0.53	0.73	0.55	0.68
UMA-M-1.1 (OMol) [20]	50M [†]	1	-	-	-	-	-	-	3.01	3.51	82.6 (0.61‡) 90.3 (0.76 [‡])	0.70	0.55	0.69
eSEN-S-OMC [22]	6M	1	1.06	5.58	0.96	1.05	5.39	0.94	3.38	5.58	1.04	1.15	0.76	0.58	0.72
eSEN-S-OMol [31]	6M	1	-	-	-	-	-	-	2.85	4.47	0.68	0.83	0.73	0.57	0.72

[†] Reported is the number of active parameters during inference, which is lower than the total number of parameters used to train UMA models [20].

‡ Reported is the result excluding five outlier systems: GLYCIN, OBEQOD, QIMKIG, QQQAUG, and UJIRIO, each with > 400 kcal/mol energy error

Comparison of OMC25 and OMol25 models

To highlight the importance of crystal-specific data, we evaluated identical MLIP architectures trained on the OMC25 dataset and the molecular dataset OMol25 [31], which contains over 100 million structures with properties computed at a higher DFT level of theory (wB97M-V/def2-TZVPD [77–79]).

The relative performance of models trained on OMC25 and OMol25 depends on the evaluation task. For the X23b benchmark [73], which involves predicting lattice energies and volumes of molecular crystals, most UMA models achieve comparable accuracy in energy predictions, with OMol25 models outperforming in volume predictions. This suggests that molecular-level data can be sufficient for predicting certain properties of crystals of small to middle-sized molecules. In contrast, for the Schrödinger polymorph ranking task [74], which requires precise energy ranking of closely related crystal structures, OMC25 models outperform OMol25 models in energy metrics, although their rank correlation scores are similar. These results underscore the critical role of explicit crystal environment data in capturing subtle intermolecular interactions and packing effects necessary for accurate polymorph evaluations. It is also important to note that the DFT levels of theory used for OMC25 and OMol25 differ from the reference level applied in the Schrödinger polymorph ranking task. When focusing on models trained solely on a single dataset (eSEN models), OMol25-trained models perform better; however, when comparing models trained on both datasets (UMA models), those with the OMC task show superior performance. Overall, the findings demonstrate the complementary strengths of

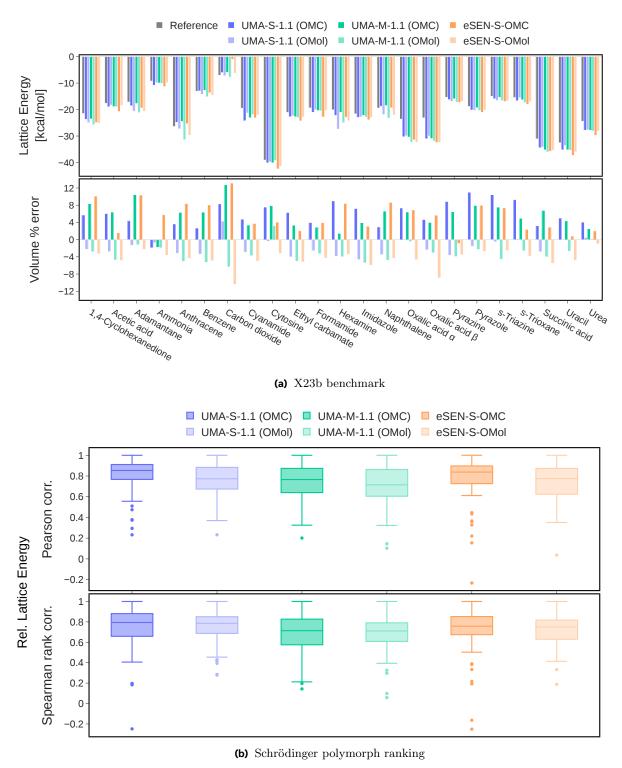


Figure S2 MLIP evaluation results for energy-conserving models: **(a)** per system values for the lattice energies and volume percentage errors for the X23b benchmark systems, **(b)** distributions of Pearson and Spearman rank correlations of relative lattice energies for all systems considered in Schrödinger polymorph ranking.

molecular and crystal datasets depending on the specific objectives.

Model Hyperparameters

Table S2 summarizes the model and training parameters for the eSEN [22] and EquiformerV2 [29] models trained on the OMC25 dataset. The eSEN model was trained in two stages: first, a direct model with a maximum of 30 neighbors and without a stress head was trained, and, subsequently, an energy-conserving model with up to 300 neighbors and an additional stress loss was trained. Detailed descriptions of the UMA models are provided in [20], and the eSEN-S-OMol model is described in [31].

Table S2 Hyperparameters and training details for the eSEN [22] and EquiformerV2 [29] models trained on the OMC25 dataset.

Hyperparameters	eSEN-S-OMC	eqV2-S-OMC
Number of parameters	6M	31M
Maximum number of neighbors	30, 300	30
Cutoff radius (Å)	6	6 & 12
Number of layers	4	8
Number of sphere channels	128	128
Number of edge channels	128	128
Maximum degree L_{max}	2	4
Maximum order M_{max}	2	2
Distance function	gaussian	gaussian
Number of distance basis	64	512
Number of hidden channels	128	-
Normalization type	rms_norm_sh	layer_norm_sh
Activation type	gate	-
ff_type	spectral	-
Number of Transformer blocks	-	8
Dimension of hidden scalar features in radial functions d_{edge}	-	(0, 128)
Embedding dimension d_{embed}	-	(4, 128)
$f_{ij}^{(L)}$ dimension d_{attn_hidden}	-	(4, 64)
Number of attention heads h	-	8
$f_{ij}^{(0)}$ dimension d_{attn_alpha}	_	(0, 64)
Value dimension d_{attn_value}	_	(4, 16)
Hidden dimension in feed forward networks d_{ffn}	_	(4, 128)
Grid resolution R	-	18
Number of GPUs	32	64
Optimizer	AdamW	AdamW
Learning rate scheduling	Cosine	Cosine
Warmup epochs	0.1	0.01
Warmup factor	0.2	0.2
Maximum learning rate	8×10^{-4}	6×10^{-4}
Minimum learning rate factor	0.01	0.01
Gradient clipping norm threshold	100	100
Model EMA decay	0.999	0.999
Weight decay	1×10^{-3}	1×10^{-3}
Dropout rate	-	0.1
Batch size	10,016 atoms	76,800 systems
Number of epochs	4, 2.4	150
Stochastic depth	-	0.1
Energy loss coefficient	10, 10	10
Force loss coefficient	30, 2	5
Stress loss coefficient	0, 1	1