MatSciML: A Broad, Multi-Task Benchmark for Solid-State Materials Modeling

Kin Long Kelvin Lee
Intel AXG

Carmelo Gonzales
Intel Labs

Marcel Nassar
Intel Labs

Matthew Spellings
Vector Institute

Mikhail Galkin
Intel Labs

Santiago Miret*
Intel Labs

Abstract

We propose MatSci ML, a novel benchmark for modeling **Mat**erials **Sci**ence using **M**achine **L**earning methods focused on solid-state materials with periodic crystal structures. Applying machine learning methods to solid-state materials is a nascent field with substantial fragmentation largely driven by the great variety of datasets used to develop machine learning models. This fragmentation makes comparing the performance and generalizability of different methods difficult, thereby hindering overall research progress in the field. Building on top of open-source datasets, including large-scale datasets like the OpenCatalyst, OQMD, NOMAD, the Carolina Materials Database, and Materials Project, the MatSci ML benchmark provides a diverse set of materials systems and properties data for model training and evaluation, including simulated energies, atomic forces, material bandgaps, as well as classification data for crystal symmetries via space groups. The diversity of properties in MatSci ML makes the implementation and evaluation of multi-task learning algorithms for solid-state materials possible, while the diversity of datasets facilitates the development of new, more generalized algorithms and methods across multiple datasets. In the multi-dataset learning setting, MatSci ML enables researchers to combine observations from multiple datasets to perform joint prediction of common properties, such as energy and forces. Using MatSci ML, we evaluate the performance of different graph neural networks and equivariant point cloud networks on several benchmark tasks spanning single task, multitask, and multi-data learning scenarios. Our open-source code is available at `https://github.com/IntelLabs/matsciml`.

1 Introduction

Solid-state materials provide the foundation for a diverse set of modern technologies, such as computer hardware, batteries, biomedical implants, and catalysts. Discovering, modeling, evaluating, and understanding of solid-state materials will therefore continue to play a significant role in complex technological challenges of the future, such as clean energy and transportation, sustainable agriculture, and personalized healthcare. The ability to accurately and efficiently model materials properties, as well as complex materials behavior under diverse conditions remains a major challenge in materials design. As such, machine learning (ML) methods have been increasingly applied to develop property prediction models that exhibit significantly greater computational efficiency compared to traditional physics-based methods,

*Correspondence to santiago.miret@intel.com.

AI for Accelerated Materials Design Workshop at NeurIPS 2023.

such as density functional theory (DFT) [6, 31]. Given this challenge, a variety of deep learning models and methods have been proposed to solve concrete challenges involving DFT data [20, 14, 7]. Generally, the research has focused on datasets targeted for concrete applications, such as the OpenCatalyst Dataset (OCP)[6, 43] for catalytic materials and Materials Project (MP) [25] for a broad range of solid-state materials with relevance to clean energy. Many of the aforementioned methods often focus on a distinct set of properties (e.g., energy and force prediction), which often have limited use for practical applications [16].

Given the current state-of-the-art, there is a need for more comprehensive ways to evaluate the modeling capabilities of machine learning models for solid-state materials. Evaluations should contain both a broader range of materials systems and their associated properties with the goal of enabling the design of more generalizable and versatile models. Based on the success of benchmarks inspiring research advances in computer vision [11], natural language processing [44, 41], molecular modeling [47, 5, 18, 24] and protein modeling [50] amongst other fields, we develop a benchmark for **Mat**erials **Sci**ence modeling using **M**achine **L**earning modeling (**MatSci ML**) [2] targetting periodic crystal structures. MatSci ML brings the following capabilities and features towards comprehensive solid-state materials benchmarking:

1. **Data Diversity:** MatSci ML integrates multiple open-source datasets, leading to a broader diversity of materials structures and properties covered by the benchmark as described in Section 3.
2. **Multi-Task Training:** MatSci ML includes support for multi-task training methods across multiple regression and classification targets for ML models. This enables researchers to leverage multi-task training methods for solid-state materials modeling on both graph-based and point cloud based representations as shown in Section 5.
3. **Multi-Dataset Integration:** MatSci ML enables joint training of machine learning models on heterogeneous data from different datasets in a unified manner. This facilitates and encourages research towards generalizable, efficient, and accurate ML models and methods for solid-state materials as described in Section 5.

To the best of our knowledge, MatSci ML is the first benchmark to enable multi-task and multi-dataset learning for solid-state materials. We describe related work in Section 2, introduce benchmark tasks in Section 3, formally define all learning settings in Section 4, and provide an analysis of their performance in Section 5.

2 Related Work

Research at the intersection of materials science and machine learning has been growing in recent years [32, 41, 45]. While adjacent research work in molecular modeling has seen significant increases in recent years, modeling of solid-state materials with periodic crystal structures has been comparatively underexplored.

Molecular Modeling: Applying machine learning to predict properties and design molecules has been an active area of research in recent years. This research has spanned many different dimensions including the development of benchmarks for property prediction [33, 35, 1, 15, 23, 47] and molecular design [47, 5, 18, 24, 1]. This, in turn, has facilitated the development of a diverse set of machine learning methods for molecular property prediction, many of which are based on graph neural networks and geometric deep learning models that include various types of useful inductive biases [21, 39, 19, 17, 20]. Additionally, there has also been a significant amount of research exploring graph-based molecular generation algorithms whose performance is evaluated on the aforementioned benchmarks [53, 52, 4, 40, 26, 56]. Solid-state materials differ significantly from molecules given their periodic crystal structure, which greatly affects their properties and behavior. This periodic structure creates the need for different representations and modeling methods that resolve greater degrees of symmetries and geometrical features found in solid-state materials [9].

[2] https://github.com/IntelLabs/matsciml

Solid-State Materials Modeling: Compared to molecular structures, the study of solid-state materials has seen significantly less ML research activity. While there has been some work on graph-based property prediction for solid-state materials [51, 8, 7, 27, 48, 10, 6, 31], many papers evaluate their proposed methods on different datasets making it difficult to compare their overall performance. This tendecy also holds in research work on the generation of solid-state crystal structures [49, 22, 46] where each method is evaluated according to the setting the authors propose. While there has been some work aiming to standardize the evaluation of machine learning models for property prediction [13], much of this work has been limited to small-scale datasets. Additionally, unlike for molecules where graph and text-based representations have been quite successful, descriptive and scalable representations for crystal structures remain an active area of research [9]. Overall, this creates a need for more comprehensive benchmarks for solid-state materials with large and chemically diverse datasets that enable more thorough studies of learned representations of solid-state materials, in addition to the development of new machine learning methods.

3 Benchmark Tasks

The MatSci ML benchmark comprises 10 tasks across 3 different task groups. MatSci ML leverages the Open MatSci ML Toolkit [31] as the backbone platform with basic primitives, including support for graph and point-cloud based data structures, as well as modeling capabilities to support the diverse set of tasks. Building on top of the Open MatSci ML Toolkit enables the addition of new tasks and datasets in a modular manner for desired future capabilities ranging from generative modeling to ML potentials for dynamical simulations. We outline the different task definitions, data sources, dataset statistics and evaluation metrics in Table 1 and will describe them in detail in subsequent sections.

Table 1: Benchmark task descriptions. Each task, along with its category, the source of dataset, the size of each split and evaluation metric are shown below. *Abbr.*, Reg.: regression; Class.: classification; ACC: accuracy; MSE: mean-square error; MAE: mean average error

Task	Task Category	Data Source	#Train	#Validation	#Test	Metric
Energy Prediction Tasks						
S2EF	Property Reg.	OpenCatalyst Project [6]	2,000,000	1,000,000	-	MSE
IS2RE	Property Reg.	OpenCatalyst Project [6]	500,000	25,000	-	MSE
Formation Energy	Property Reg.	Materials Project [25]	108,159	30,904	15,456	MSE
LiPS	Property Reg.	LiPS [2]	17,500	5,000	2,500	MSE
OQMD	Property Reg.	OQMD [28]	818,076	204,519	-	MSE
NOMAD	Property Reg.	NOMAD [12]	111,056	27,764	-	MSE
CMD	Property Reg.	Carolina Materials Database [55]	171,548	42,887	-	MSE
Force Prediction Tasks						
S2EF	Property Reg.	OpenCatalyst Project [6]	2,000,000[1]	1,000,000	-	MAE
LiPS	Property Reg.	LiPS [2]	17,500	5,000	2,500	MAE
Property Prediction Tasks						
Material Bandgap	Property Reg.	Materials Project [25]	108,159	30,904	15,456	MSE
Fermi Energy	Property Reg.	Materials Project [25]	108,159	30,904	15,456	MSE
Stability	Property Class.	Materials Project [25]	108,159	30,904	15,456	ACC
Space Group	Property Class.	Materials Project [25]	108,159	30,904	15,456	ACC

3.1 Energy Prediction Tasks

Energy prediction is one of the most common property prediction tasks in both molecular and solid-state crystal structure modeling, and is generally included in most relevant datasets [6, 25, 2, 33]. Energy is a critical property of a material system that indicates how stable the materials system is. Moreover, the energy can be used to understand many different aspects of materials behavior, and has also inspired methods development in machine learning, such as "energy-based learning" methods. The ubiquity of energy labels in various datasets allows us to combine multiple datasets in a *multi-data* setting. The collection of data in energy prediciton spans ∼1.5 million bulk materials from various sources and relaxation trajectory data diverse adsorbate + surface + bulk combinations from OpenCatalyst.

Structure to Energy & Forces (S2EF) from OCP [6] requires prediction of the adsorption energy of a molecular adsorbate on a catalyst surface. We directly adopt the dataset splits from OCP containing a training set, an in-distribution validation set, and a set of out-of-distribution validation sets based on different molecular absorbates or catalysts. Accurate prediction of adsorbate-surface interactions is necessary for effective materials design in many applications, including catalysts and semiconductors.

Initial Structure to Relaxed Energy (IS2RE) from OCP [6] involves the prediction of relaxed adsorption energy of a molecular adsorbate on a solid-state catalyst surface. We directly adopt the dataset splits from OCP containing a training set, an in-distribution validation set, and a set of out-of-distribution validation sets based on different molecular adsorbates or catalysts. Predicting the relaxed adsorption energy of joint molecular and solid-state materials from an initial structure has substantial impact on the design of catalytic materials which can help accelerate a variety of chemical reactions. This task can help understand the influence of solid-state material composition and structure, as well as its interactions with molecules.

Formation Energy from MP [25] involves predicting the energy of the material relative to its constitutients, as a function of the relative three-dimensional arrangement of atoms in the unit cell. MP normalizes the formation energy based on the stoichiometry of the material (e.g. H_2O, SiO_2) in units of eV/atom. We construct a dataset split of MP where the representation of different crystal structures is consistent across training, validation, and test sets. Formation energy, along with entropy, determines the thermodynamic stability of a material, and thus how feasible it is for the material to be experimentally synthesized and what applications it may be suitable for. This task could be applicable to materials design of bulk solid-state materials, as opposed to the exposed surfaces found in OCP.

LiPS Energy from the LiPS dataset [2] involves the prediction of the energy of LiPS material structures as they evolve dynamically relative to a reference configuration, in units of meV/atom. We construct a random dataset split based on the original dataset similar to prior work [16]. Reliably accurate predictions of the energy of a configuration, meaning atoms in space, are needed for ML potentials used in simulations of materials under physically relevant conditions, such as room temperature and atmospheric pressure.

OQMD from the OQMD dataset [28] involves the prediction of the formation energy of a material structure measured in eV/atom based on the DFT calculations. We construct a random dataset split based of 1,022,595 bulk material structures in the dataset with a 20% validation split. OQMD represents the largest collection of bulk material formation energy calculation, including more sample than Materials Project, NOMAD and CMD combined.

NOMAD from the NOMAD dataset [12] involves the prediction of the formation energy of a material structure measured in eV/atom based on crowdsourced calculations. We construct a random dataset split based of 138,820 bulk material structures in the dataset with a 20% validation split.

CMD from the Carolina-MatDB dataset [55] involves the prediction of the formation energy of a material structure measured in eV/atom based on the DFT calculations of structures discovered by machine learning methods. We construct a random dataset split based of 214,435 bulk material structures in the dataset with a 20% validation split.

3.2 Force Prediction Tasks

Many workflows for machine-learned potentials harness automatic differentiation available in modern ML frameworks to produce a conservative potential energy function U, linked to the force $\vec{f}$ *via* the gradient: $\vec{f} = -\nabla U$. While this conservative formulation could be important for fine-scale thermodynamic stability of simulations, for some applications learning to predict forces independently from energy—either in a rotation-equivariant or non-rotation-equivariant way—may also suffice. All models described here derive forces from the gradient of a conservative potential energy.

Structure to Energy & Forces (S2EF) from OCP [6] includes both energy and force labels; the latter represents the force exerted on each atom within the molecular adsorbate in units of eV/Å. We adopt the same dataset splits found in OCP. Predicting forces on each atom for a snapshot of particle configurations is needed for structure relaxation and other optimization methods used to find low-energy states of materials systems. Accurate force predictors also provide concrete opportunities to incorporate machine learning models into classical materials modeling workflows such as molecular dynamics simulations [2, 7, 16].

LiPS Forces from the LiPS dataset [2] includes per-atom forces (in meV/Å) based on a random split. Similar to S2EF, predicting the atomic forces of a system in a generalizable way would enable applying machine learning to further understand materials behavior.In contrast to S2EF, this dataset comprises many frames of a single Li-ion system, as opposed to a diverse set of compositions and structures.

3.3 Property Prediction Tasks

For all Materials Project (MP) [25] property prediction tasks in this section, we apply the same dataset split as for the formation energy described above. In this case, the representation of different crystal structures is consistent across training, validation, and test sets.

Material Bandgap involves the prediction of the bandgap of a solid-state material in units of eV, corresponding to the amount of energy required to promote a valence electron into the conduction band. Larger bandgaps imply low electronic conducitivty of the material (e.g. insulators), while small bandgap imply large electronic conductivity (e.g. metals) with many materials being somewhat conductive (e.g. semiconductors). Predicting the bandgap of a material is critical for many electronic materials and their applications, such as semiconductors for computer hardware and photovoltaics. This task aims to understand how the design (e.g. composition and configuration) of crystal structures affects the bandgap.

Fermi Energy is the highest occupied energy level of a material at absolute zero temperature measured in eV, which correlates with the conductivity of a material. The Fermi energy generally represents the halfway point between the valence and conduction band and is thereby closely related to the material bandgap. Predicting the Fermi energy can help understand the electric properties of a given material, which can in turn be used to engineer the conductivity characteristics of materials for new applications.

Stability is a binary classification task to predict whether a given material configuration is thermodynamically stable at absolute zero. Understanding material stability is particularly relevant for evaluating and conditioning generative models, e.g., preferentially sampling from stable configurations of chemical space should result in experimentally viable materials.

Space Group is a multiclass classification task to predict which, of the 230 possible crystallographic space groups, a given material belongs to. Predicting the space group requires embedding the effect of symmetry operations (e.g. rotation and exchange) of a solid-state structure, which ultimately influences its physical properties and stability.

4 Training Methods

We apply a set of deep learning models and training methods to showcase the capabilities of the benchmark and derive some interesting insights. While we believe these baselines are representative of the general capabilities of deep learning methods for materials modeling, our experiments are unlikely to achieve the best possible modeling performance. As such, we encourage future work to leverage the benchmark to improve upon currently available methods, as well as further research into the development of new methods. First, we describe the different deep learning architectures we studied.

Graph Neural Networks (GNNs) encode the material structure as a graph where the atoms generally represent the nodes and the edges are the connections between the atoms.

Unlike molecular structures, solid-state materials do not have a canonical way to encode bonds between different atoms. As such, distance based radius graphs are used to construct the graph of the corresponding material. We apply MegNet [8] across all tasks in MatSci ML to understand the performance of domain-specific graph neural networks.

Equivariant Graph Neural Networks encode rotational equivariance into their architecture, which is a useful inductive bias for materials property prediction. Regular GNNs do not have rotational equivariance or scalar invariance by default in their architecture and have to be intentionally encoded. We apply E(n)-GNN [38] across all tasks in MatSci ML to understand the performance of equivariant graph neural networks.

Short-Range Equivariant Models operate on a point cloud data structure where local neighborhoods in the point cloud receive the greatest importance in parameter updates of the neural network. The additional flexibility of the point cloud data structure also helps promote localized representations of relevant elements in the materials structure through targeted mathematical formulations, such as Clifford algebras [42, 37, 36], which facilitate efficient model training. We apply GALA [42] across all tasks in MatSci ML to understand the performance of short-range equivariant networks.

4.1 Single Task vs Multi-Task Learning

Throughout this paper, we refer to a "task" a mapping from a given set of inputs (which may come from a specific dataset) to a desired single outcome (e.g., classification, regression) encapsulated by a single loss function. *Multi-task training* refers to training a model on more than one type of loss function, such as regression *and* classification jointly performed on the MP dataset. *Multi-data training* refers to training a model on a similar type of label across multiple data sources, such as energy prediction on diverse materials drawn jointly from the OCP dataset and MP dataset. Next, we outline formal definitions of the three methods.

Single Task Learning is a common way to approach solid-state materials modeling by training a model exclusively on one task at a time. In this case, the model learns a mapping function (f) between input (x) and output (y) where $x \in t_n \in \mathcal{T}$ is drawn from a pool of tasks $\mathcal{T}$. The learning objective is summarized by a single loss $\mathcal{L}_t$ that is minimized for may include multiple regression *or* classification targets from the same dataset.

Multi-Task Learning aims to learn a mapping function (f) between input (x) and output (y) from different tasks t_n, i.e. $x = [x_{t_1}, x_{t_2}, ...x_{t_n}]$ and $y = [y_{t_1}, y_{t_2}, ...y_{t_n}]$. In this paper, we study multi-task learning using a joint encoder with a predetermined model architecture followed by task-specific output heads. To remain within reasonable compute budgets, we perform experiments on two tasks at a given time with a balanced loss between the two tasks: $\mathcal{L}_\theta = \mathcal{L}_{t_1} + \mathcal{L}_{t_2}$. In this setting, both losses backpropagate gradients to the joint encoder in addition to their respective output heads. Additionally, we perform multi-task learning using PCGrad [54] which aims to minimize gradient conflicts between different tasks.

Multi-Data Learning aims to learn a mapping function (f) between input (x) and output (y) from different datasets d_n, i.e. $x = [x_{d_1}, x_{d_2}, ...x_{d_n}]$ and $y = [y_{d_1}, y_{d_2}, ...y_{d_n}]$. In this case, the output (y) is a single property found among each of the datasets, such as a measurement of energy or atomic forces. Similar to multi-task learning, we study multi-data learning using a joint encoder with a predetermined model architecture followed by task-specific output heads. We perform experiments on two datasets at a time with a balanced loss between the two tasks: $\mathcal{L}_\theta = \mathcal{L}_{d_1} + \mathcal{L}_{d_2}$.

5 Experiments

We perform various experiments across the different models and methods described in Section 4, including training all models for single-task and multi-task learning shown in Table 2 and Table 3, respectively, as well as multi-data learning for E(n)-GNN and MegNet shown in Table 4. We did not perform multi-data learning for GALA given the increased computational cost of training the model on the large combined dataset, especially S2EF and IS2RE, compared to the other methods. Our general results also indicate that GALA

underperforms compared to other models, suggesting that it would be more productive to focus multi-data experiments on E(n)-GNN and MegNet.

5.1 Single-Task Learning

We perform single task learning for all tasks in MatSci ML with the results summarized in Table 2. For additional reference, we add state-of-the-art results for OpenCatalyst OC-20 data based on the public OC20 leaderboard. The results on the leaderboard represent test data splits that are only available through the leaderboard interface, while our results are based on the publicly available validation data splits. For both Materials Project (MP) and LiPS, we create new dataset splits which make it difficult to compare to existing results reported in the literature. The results from Table 2 indicate that:

Table 2: Benchmark results on single-task learning. We report the validation set performance for each experiment and highlight the best performance among all models; SOTA model performance from literature is added where applicable; "-" indicates a non-applicable setting. Graph-based models perform better than point cloud based models on single task learning.

Task	Metric	Equivariant Neural Network	Graph Neural Network	Point Cloud Network	Literature SOTA
		E(n)-GNN	MegNet	GALA	
Energy Prediction					
S2EF	MSE	0.826	1.252	6.611	0.227 (Equiformer [30])
IS2RE	MSE	0.186	0.229	5.133	0.300 (Equiformer [30])
MP	MSE	0.045	0.100	0.32	-
LiPS	MSE	0.579	0.989	0.985	-
OQMD	MSE	0.244	0.276	-	-
NOMAD	MSE	0.209	0.215	-	-
CMD	MSE	0.029	0.141	-	-
Force Prediction					
S2EF	MAE	0.957	0.186	567.4	0.0138 (Equiformer [30])
LiPS	MAE	0.443	0.443	1.078	-
Property Prediction					
Band	MSE	0.504	0.497	1.234	-
Fermi	MSE	0.859	0.849	3.506	-
Stable	ACC	79.9	83	77.2	-
Space	ACC	29.8	31.3	20.1	-

Graph neural networks perform well across all tasks. E(n)-GNN outperforms all other models across the energy prediction tasks, while MegNet performs best for force prediction and MP-based tasks. Both E(n)-GNN and MegNet outperform GALA across all tasks in MatSci ML . This suggests that graph-based data structures provide a useful inductive bias for modeling solid-state materials although a more thorough study is required to further confirm this observation. The reported results from the OC20 leaderboard indicate that the evaluated models are far from SOTA performance in S2EF, both for energy and forces, but may be competitive for IS2RE.

Space group classification is a difficult task for all models. All evaluated models perform poorly on space group classification with MegNet reaching an accuracy of 31.3%. The difficulty associated with this task is twofold: the natural imbalance of class labels owing to the fact that materials of certain space groups are more prevalent than others, and that symmetry operations are hierarchical, thus requiring models to differentiate between groups with similar bases. The latter reinforces prior findings that models and representations which specifically include higher-order symmetry could be useful for solid-state materials [27].

5.2 Multi-Task Learning

We then probe the multi-task learning scenario based on property prediction tasks from Materials Project data splits spanning both regression and classification. MP provides the greatest diversity of labels for evaluating different property prediction targets suitable for the multi-task setting. We study the multi-task performance under the settings described in Section 4.1 with additive task losses for joint backpropagation, as well as for PCGrad [54]. Based on the results in Table 3, we observe:

Table 3: Benchmark results on materials project multi-task learning. We show the best performing along with single-task baseline with each multi-task run outperforming the single-task baseline also highlighted. Multi-task learning generally outperforms single task on regression tasks with only small performance difference between additive losses and PCGrad.

Task	Metric	E(n)-GNN				MegNet				GALA			
		+Band	+Fermi	+Stable	+SG	+Band	+Fermi	+Stable	+SG	+Band	+Fermi	+Stable	+SG
Multitask Training (Additive Losses)													
Band	MSE	0.504	0.389	0.314	0.43	0.497	0.454	0.368	0.585	1.23	0.622	0.51	0.54
Fermi	MSE	0.211	0.859	0.25	0.499	0.284	0.849	0.263	0.648	0.606	3.51	0.508	0.676
Stable	ACC	81.6	77.4	79.9	78.1	80.6	76.9	83.0	76.9	78.0	77.0	77.2	76.9
Space	ACC	30.6	27.5	30.9	29.8	30.8	31.1	21.6	31.3	18.2	18.4	19.9	20.1
PCGrad Training [54]													
Band	MSE	0.504	0.389	0.312	0.406	0.497	0.454	0.343	0.537	1.23	0.622	0.511	0.563
Fermi	MSE	0.211	0.859	0.259	0.43	0.284	0.849	0.268	0.452	0.606	3.506	0.461	0.622
Stable	ACC	81.9	77.7	79.9	77.7	81.3	76.9	83.0	77.4	77.0	77.1	77.2	77.7
Space	ACC	30.3	27.5	30.1	29.8	23.8	25.4	26.6	31.3	19.1	22.2	23.2	20.1

Multi-task learning generally improves task performance on individual tasks. Task performance in the multi-task setting generally improves across all of the tasks studied. This is particularly true for the regression tasks (bandgap and fermi energy) and less so for the classification tasks where performance remains similar to single-task learning. This suggests that many of the tasks in MP have a high degree of correlation leading to overall better learning.

PCGrad offers small improvements in multi-task learning. The results across all three models studied indicate that PCGrad provides little performance improvement compared to multi-task learning with additive losses. This further reinforces the idea that the tasks in MP have a high degree of correlation given that one of the primary goals of PCGrad is to resolve gradient conflicts between different tasks. Hence, a low degree of gradient conflicts in highly correlated tasks leads to only small performance gains.

5.3 Multi-Data Learning

We perform multi-data learning for energy and force prediction across all different datasets. Based on the results shown in Table 4, we observe:

IS2RE energy performance worsens with multi-data learning. IS2RE energy worsens in the multi-data setting for both E(n)-GNN and MegNet. We hypothesize that is due to the fact that IS2RE aims to predict relaxed energy of a given structure, which is different from the single frame prediction present in all other datasets.

S2EF energy performance improves with multi-data learning. S2EF energy prediction generally improves in the multi-data setting for both E(n)-GNN and MegNet with the exception of E(n)-GNN S2EF + IS2RE. This reinforces the notion that S2EF energy prediction is naturally more correlated with the energy labels in MP and LiPS given that all datasets evaluate energy at the given frame, as opposed to IS2RE which evaluates energy for a final relaxed state—methods akin to Δ-ML [34] may be required to bridge this gap.

MP and LiPS see varied results in multi-data learning. LiPS energy performance remains relatively stable for MegNet compared to the single-task performance and worsens for E(n)-GNN. MP energy prediction generally shows improvement when combined with S2EF and deterioration when combined with LiPS. This generally indicates that MP and LiPS are not very well correlated. MP improvements for MegNet in combination with S2EF and IS2RE may indicate that model is able to acquire more generalized knowledge on the larger datasets, which would have to be confirmed with more thorough studies.

Force prediction improves in multi-data learning. The improvements in force prediction between S2EF and LiPS further indicate a strong correlation between the tasks, which is also observed in energy prediction.

Table 4: Benchmark results on energy+forces multi-dataset learning. We show the best performing along with single-task baseline with each multi-task run outperforming the single-task baseline also highlighted. Multi-data outperforms the single-task baseline in some cases for both models. "-" indicates not applicable for this setting.

Task	Metric	E(n)-GNN				MegNet			
		+S2EF	+IS2RE	+MP	+LiPS	+S2EF	+IS2RE	+MP	+LiPS
Energy Prediction									
S2EF	MSE	0.826	0.282	0.744	0.193	1.252	0.455	0.376	0.445
IS2RE	MSE	0.252	0.186	0.32	0.287	0.34	0.229	0.374	0.276
MP	MSE	0.044	0.064	0.045	0.385	0.077	0.086	0.100	1.038
LiPS	MSE	0.966	0.992	0.988	0.579	0.966	0.997	0.988	0.989
Force Prediction									
S2EF	MAE	0.957	-	-	0.185	0.186	-	-	0.177
LiPS	MAE	0.361	-	-	0.443	0.441	-	-	0.443

6 Material Generation Pipeline

In addition to efficiently evaluating diverse sets of materials properties as described in Section 3, generative methods provide the opportunity to significnatly expand the space of known solid-state materials. Currently, there are only $\sim 200k$ experimentally known inorganic materials in the ICSD database [3], which is significantly smaller than the space of possible materials designs creating an vast opportunity for materials discovery.

To faciliate generative modeling in MatSciml, we applied our Materials Project dataset (described in Appendix B.2) on the generative modeling task using CDVAE [49], a latent diffusion model that trains a VAE on the reconstruction objective with DimeNet++ [19] as an encoder and GemNet-dT [20] as a decoder on the denoising objective. For the sake of numerical stability, we trained and generated samples with 25 or less atoms in the structure that resulted in 64,251 training data points, 18,142 for validation, and 9,098 for testing (denoting this subset as `mp25`). Following the standard hyperparameters reported in Xie et al. [49] (with the only change being a larger decoder cutoff radius of 12Å to account for larger structures than those in the original datasets), we trained a 5M parameter CDVAE model and sampled 10,000 structures using Langevin dynamics. The results are shown in Table 5.

Table 5: Generation quality metrics of CDVAE matching the quality of the original implementation in Xie et al. [49] with a new subsample of Materials Project (`mp25`).

Dataset	Validity (%)		Coverage (%)	
	Structure	Composition	Recall	Precision
`mp25`	99.74	89.01	97.74	99.58

The results presented in Table 5 expand upon the results in Xie et al. [49] given that original implementation only trained on a subset of 20k datapoints from Materials Project.

7 Conclusion

To the best of our knowledge, MatSci ML is the first benchmark to enable multi-task and multi-dataset learning for solid-state materials, thereby facilitating machine learning researchers to build more generalizable models to accelerate the deployment of machine learning tools in the design, discovery, and evaluation of new materials systems. The focus on multi-task models can enable future work on *generalist* models that can be applied to various downstream applications in the materials science domain. These models could follow pre-training procedure on auxillary tasks, such as symmetry classification [29], for general knowledge acquisition followed by specialized fine-tuning for a given task on limited data.

References

[1] Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1): 185, 2022.

[2] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

[3] Alec Belsky, Mariette Hellenbrandt, Vicky Lynn Karen, and Peter Luksch. New developments in the inorganic crystal structure database (icsd): accessibility in support of materials research and design. *Acta Crystallographica Section B: Structural Science*, 58(3):364–369, 2002.

[4] Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, 34:27381–27394, 2021.

[5] Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3):1096–1108, 2019.

[6] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10): 6059–6072, 2021.

[7] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

[8] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019.

[9] James Damewood, Jessica Karaguesian, Jaclyn R Lunger, Aik Rui Tan, Mingrou Xie, Jiayu Peng, and Rafael Gómez-Bombarelli. Representations of materials for machine learning. *Annual Review of Materials Research*, 53, 2023.

[10] Pierre-Paul De Breuck, Geoffroy Hautier, and Gian-Marco Rignanese. Materials property prediction for limited datasets enabled by feature selection and joint learning with modnet. *npj Computational Materials*, 7(1):83, 2021.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[12] Claudia Draxl and Matthias Scheffler. The nomad laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials*, 2(3):036001, 2019.

[13] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020.

[14] Alexandre Duval, Victor Schmidt, Alex Hernandez Garcia, Santiago Miret, Fragkiskos D Malliaros, Yoshua Bengio, and David Rolnick. Faenet: Frame averaging equivariant gnn for materials modeling. In *International Conference on Machine Learning*, 2023.

[15] Peter Eastman, Pavan Kumar Behara, David L Dotson, Raimondas Galvelis, John E Herr, Josh T Horton, Yuezhi Mao, John D Chodera, Benjamin P Pritchard, Yuanqing Wang, et al. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(1):11, 2023.

[16] Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

[17] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33:1970–1981, 2020.

[18] Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor Coley. Sample efficiency matters: a benchmark for practical molecular optimization. *Advances in Neural Information Processing Systems*, 35:21342–21357, 2022.

[19] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020.

[20] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34:6790–6802, 2021.

[21] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[22] Prashant Govindarajan, Santiago Miret, Jarrid Rector-Brooks, Mariano Phielipp, Janarthanan Rajendran, and Sarath Chandar. Behavioral cloning for crystal design. In *Workshop on"Machine Learning for Materials"ICLR 2023*, 2023.

[23] Johannes Hoja, Leonardo Medrano Sandonas, Brian G Ernst, Alvaro Vazquez-Mayagoitia, Robert A DiStasio Jr, and Alexandre Tkatchenko. Qm7-x, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Scientific data*, 8(1):43, 2021.

[24] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[25] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1):011002, 2013.

[26] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures. In *International conference on machine learning*, pages 4849–4859. PMLR, 2020.

[27] Oumar Kaba and Siamak Ravanbakhsh. Equivariant networks for crystal structures. *Advances in Neural Information Processing Systems*, 35:4150–4164, 2022.

[28] Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Computational Materials*, 1(1):1–15, 2015.

[29] Kin Long Kelvin Lee, Carmelo Gonzales, Matthew Spellings, Mikhail Galkin, Santiago Miret, and Nalini Kumar. Towards foundation models for materials science: The open matsci ml toolkit. In *Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis*, SC-W '23, page 51–59, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400707858. doi: 10.1145/3624062.3626081. URL `https://doi.org/10.1145/3624062.3626081`.

[30] Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. *arXiv preprint arXiv:2206.11990*, 2022.

[31] Santiago Miret, Kin Long Kelvin Lee, Carmelo Gonzales, Marcel Nassar, and Matthew Spellings. The open matsci ml toolkit: A flexible framework for machine learning in materials science. *arXiv preprint arXiv:2210.17484*, 2022.

[32] Santiago Miret, Marta Skreta, Benjamin Sanchez-Lengelin, Shyue Ping Ong, Zamyla Morgan-Chan, and Alan Aspuru-Guzik. Ai4mat: Ai for accelerated materials design neurips 2022 workshop, 2022. URL `https://sites.google.com/view/ai4mat`.

[33] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1, 2014.

[34] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096, May 2015. ISSN 1549-9618. doi: 10.1021/acs.jctc.5b00099.

[35] Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

[36] David Ruhe, Johannes Brandstetter, and Patrick Forré. Clifford group equivariant neural networks, 2023.

[37] David Ruhe, Jayesh K. Gupta, Steven de Keninck, Max Welling, and Johannes Brandstetter. Geometric clifford algebra networks. In *International Conference on Machine Learning*, 2023.

[38] Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.

[39] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

[40] Gregor Simm, Robert Pinsler, and José Miguel Hernández-Lobato. Reinforcement learning for molecular design guided by quantum mechanics. In *International Conference on Machine Learning*, pages 8959–8969. PMLR, 2020.

[41] Yu Song, Santiago Miret, and Bang Liu. Matsci-nlp: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[42] Matthew Spellings. Geometric algebra attention networks for small point clouds. *arXiv preprint arXiv:2110.02393*, 2021.

[43] Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, et al. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, 2023.

[44] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2018.

[45] Hanchen Wang, Simon Batzner, Tian Xie, and Xiang Fu. Ml4materials from molecules to materials, 2023. URL `https://www.ml4materials.com/`.

[46] Shiyu Wang, Xiaojie Guo, and Liang Zhao. Deep generative model for periodic graphs. In *Advances in Neural Information Processing Systems*, 2022.

[47] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[48] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120 (14):145301, 2018.

[49] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations*, 2022.

[50] Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. PEER: a comprehensive and multi-task benchmark for protein sequence understanding. *Advances in Neural Information Processing Systems*, 35:35156–35173, 2022.

[51] Takenori Yamamoto. Crystal graph neural networks for data mining in materials science. *Research Institute for Mathematical and Computational Sciences, LLC*, 2019.

[52] Soojung Yang, Doyeong Hwang, Seul Lee, Seongok Ryu, and Sung Ju Hwang. Hit and lead discovery with explorative rl and fragment-based molecule generation. *Advances in Neural Information Processing Systems*, 34:7924–7936, 2021.

[53] Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in neural information processing systems*, 31, 2018.

[54] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020.

[55] Yong Zhao, Mohammed Al-Fahdi, Ming Hu, Edirisuriya MD Siriwardane, Yuqi Song, Alireza Nasiri, and Jianjun Hu. High-throughput discovery of novel cubic crystal materials using deep generative neural networks. *Advanced Science*, 8(20):2100566, 2021.

[56] Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):1–10, 2019.

A Experiment Descriptions

A.1 Compute Details

We used GPU nodes on an internal cluster where a single node generally consists 8 GPU's of either: Nvidia Titan V, Titan X, or Titan Xp. Single task experiments are trained on one node for a minimum of 20 epochs, and a maximum of 50 epochs. In the case of single task training, we also apply early stopping with a patience of 15 epochs. We train multi-data experiments on one node for a maximum of 25 epochs and multi-task experiments on a single GPU for a maximum of 50 epochs.

For experiments involving OpenCatalyst tasks, we we rely on the original training and validation dataset splits and construct our own dataset splits for Materials Project and LiPS. We discuss all relevant details for dataset license and split construction in Appendix B. All experiments are conducted with the dataset split described in Table 1 with the exception of S2EF where we perform single task training on S2EF with 2M training samples and multi-data with 200k training samples to better manage the compute cost and dataset balance.

A.2 Hyperparameters

We outline the hyperparameters for all three models described in Section 5. We maintained consistent architecture parameters for all training settings across all tasks. Full set of training and evaluation parameters will be published with code release upon publication.

Table 6: Hyperparameters for E(n)-GNN

Hyperparameter	Value
MLP hidden dim	32
MLP output dim	128
# of EGNN layers	5
Node MLP dim	$[128, 128, 128]$
Edge MLP dim	$[128, 128, 128]$
Atom position MLP dim	$[64, 64]$
MLP activation	ReLU
Graph read out	Sum
Node projection block depth	3
Node projection hidden dim	128
Node projection activation	ReLU
Output block depth	3
Output hiddem dim	128
Output activation	ReLU
Optimizer Parameters	
Learning Rate	0.0001
Gamma	0
Batch Size	32

Table 7: Hyperparameters for MegNet

Hyperparameter	Value
Edge MLP dim	2
Node MLP dim	5
Graph variable MLP dim	9
MegNet blocks	4
MLP hidden dims	$[128, 64]$
MegNet convolution dims	$[128, 128, 64]$
# of S2S layers	5
# of S2S iterations	4
Output projection dims	$[64, 16]$
Dropout	0.1
Optimizer Parameters	
Learning Rate	0.0001
Batch Size	32

Table 8: Hyperparameters for GALA

Hyperparameter	Value
Input dimension	200
Hidden dimension	128
Merge function	concat
Join function	concat
Rotation-invariant mode	full
Rotation-covariant mode	full
Rotation-invariant value norm	momentum
Rotation-equivariant value norm	momentum layer
Value function normalization	layer
Score function normalization	layer
Block-level normalization	layer
Optimizer Parameters	
Learning Rate	0.0001
Gamma	0
Batch Size	16

B Dataset Descriptions

B.1 OpenCatalyst

The OpenCatalyst dataset [6] was originally published with a Creative Commons Attribution 4.0 (CC BY 4.0) license. Our work leverages the implementation of the Open MatSci ML Toolkit [31] with the same license and preprocesses the S2EF and IS2RE datasets according to the instructions documented on the Open MatSci ML Github site for S2EF [3] and Zenodo release for IS2RE [4].

B.2 Materials Project

The Materials Project (MP) [25] is also released under a CC BY 4.0 license. Setting up MP datasets first requires access to the Materials Project API by creating an account

[3] `https://github.com/IntelLabs/matsciml`

[4] `https://zenodo.org/record/7411133`

on the orginal website [5]. The API key may then be set to an environment variable: `$ export MP_API_KEY=your-api-key` to interact with the command-line interface to query for specific data, or rely on pre-configured YAML configurations to process pre-defined splits we refer to in this paper.

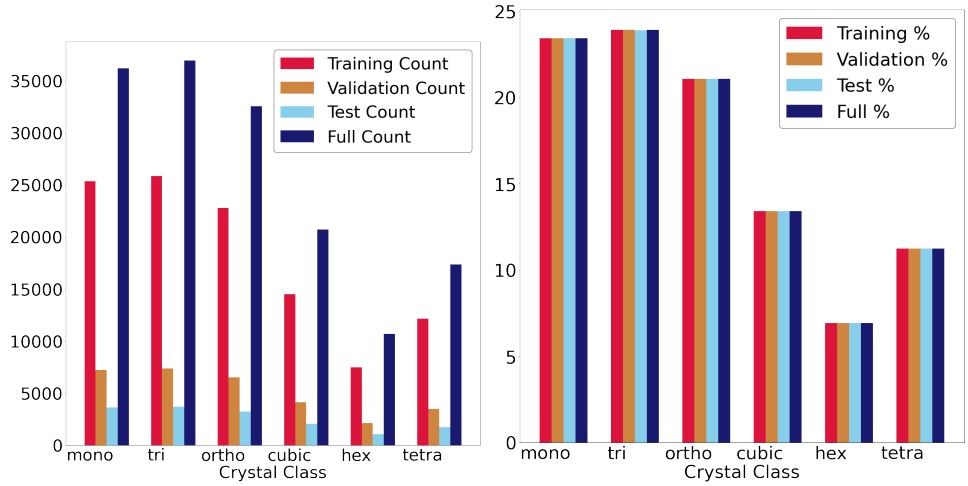

Figure 1: Dataset split of Materials Project [25] that ensures crystal structure representation across training, validation and testing splits for randomly sampled materials from the full dataset. Left panel shows data counts, while the right shows fractional composition—each split comprises the same balance in symmetry. *Abbr.*, mono: monoclinic; tri: triclinic; ortho: orthorhombic; hex: hexagonal; tetra: tetragonal

Train, validation, and test splits are defined by material id based on the split described in Figure 1, and may be found in our code in `ocpmodels/datasets/materials_project/{train, val, test}.yml`. We aimed to create a chemically balanced partitioning of the available data. We note that effective dataset splitting remains an open question without clear consensus in the broader materials community, and that past literature have generally performed custom dataset splits based on different properties of interest. As seen in the right panel of Figure 1, we partitioned the splits to preserve uniformity in the crystal family labels. Our dataset splits were informed by the fact that crystal symmetry is a universal property for all of solid-state materials that significantly affects physical properties, including structure, stability, and functional properties (e.g. band gap, magnetism). In terms of implementation, a simple command line script is used to load material id numbers and download the relevant data to lmdb files, consistent with other datasets used in Open MatSci ML Toolkit. The primary labels used for experiments includes the fields: `band_gap`, `structure`, `formula_pretty`, `efermi`, `symmetry`, `is_metal`, `is_magnetic`, `is_stable`, `formation_energy_per_atom`, `uncorrected_energy_per_atom`, and `energy_per_atom`.

To download and extract the train, validation and test datasets using our code, the following command can be used:

```
python -m ocpmodels.datasets.materials_project.cli \
    -d mp_data \
    -t base \
    -s ocpmodels/datasets/materials_project/train.yml \
    ocpmodels/datasets/materials_project/val.yml \
    ocpmodels/datasets/materials_project/test.yml
```

The `-d` flag is used to specify a directory to store the data, and defaults to `mp_data`. After running the script, the data directory will include train, validation and test folders containing

[5] `https://materialsproject.org`

lmdb files with 108159, 30904, and 15,456 samples respectively. Specifying the `-t` flag will ensure all of the main data fields listed above are included in the download.

A devset (development dataset) is also included which has 200 material samples containing the `band_gap`, and `structure` fields, which is accessible in `ocpmodels/datasets/materials_project/devset`.

Other property fields, material id's, and Materials Project's API arguments may be used with the download script to create custom datasets. Additional details on how to use the script may be found in `ocpmodels/datasets/materials_project/cli.py`.

B.3 LiPS

The LiPS dataset is also released under a CC BY 4.0 license, which can be accessed via the original release in Materials Cloud[6].

The LiPS data splits used in the experiments are included in the codebase folders `ocpmodels/datastes/lips/base/{train, val, test}`. To create the splits, we download the dataset from it's original release and split randomly into 70%, 20% and 10% chunks for training, validation and testing. A dev set is also included in `ocpmodels/datasets/lips/devset` which holds 200 samples.

C Limitations

The currently published datasets focus primarily on ground-state energy calculations at zero-temperature and pressure that include minimal information about how the material system behaves under different conditions. While OpenCatalyst includes relaxation trajectories for S2EF, they are still calculated under ideal conditions, creating the risk that behavior of the materials will be different under real-world conditions, such as room temperature and pressure. LiPS is the only dataset that includes more realistic information about material dynamics, but is limited in dataset size. Additionally, the benchmark covers only a sample relevant combinations of material properties and dataset splits available across all of the different tasks available. We hope that future work can provide more insight into how to conduct effective, potentially physics-informed dataset creation and splitting, as well as how different models can generalize across different prediction tasks. Future work is also needed to assess how to combine different types of datasets into multitask multi-data learning scenarios, which may include material types and simulation conditions.

The application of machine learning to materials science could have unintentional consequences for data privacy, where sensitive materials data is inadvertently included in a model's implicit knowledge. Similar to adjacent machine learning fields where privacy is important, future work is needed to effectively manage these situations.

[6] `https://archive.materialscloud.org/record/2022.45`