

LEARNING EQUIVARIANT MODELS BY DISCOVERING SYMMETRIES WITH LEARNABLE AUGMENTATIONS

Anonymous authors

Paper under double-blind review

ABSTRACT

Recently, a trend has emerged that favors shifting away from designing constrained equivariant architectures for data in geometric domains and instead (1) modifying the training protocol, e.g., with a specific loss and data augmentations (soft equivariance), or (2) ignoring equivariance and inferring it only implicitly. However, both options have limitations, e.g., soft equivariance still requires a priori knowledge about the underlying symmetries, while implicitly learning equivariance from data lacks interpretability. To address these limitations, we propose SEMoLA, an end-to-end approach that jointly (1) discovers a priori unknown symmetries in the data via learnable data augmentations, and uses them to (2) encode the respective approximate equivariance into arbitrary unconstrained models. Hence, it enables learning equivariant models that do not need prior knowledge about symmetries, offer interpretability, and maintain robustness to distribution shifts. Empirically, we demonstrate the ability of SEMoLA to robustly discover relevant symmetries while achieving high prediction performance across various datasets, encompassing multiple data modalities and underlying symmetry groups.

1 INTRODUCTION

In recent years, equivariant machine learning models have consolidated themselves as a powerful approach for working with geometric data, boasting increased data efficiency and generalization capabilities (Dehmamy et al., 2021; Cohen and Welling, 2016; Bulusu et al., 2021). In particular, they have shown great promise in their application to scientific domains, e.g., dynamical systems (Han et al., 2022; Xu et al., 2024), chemistry (Satorras et al., 2021; Brandstetter et al., 2021), or structural biology (Jumper et al., 2021), where the strong inductive biases imposed by the equivariant design of the models can lead to improved performance and scientific validity.

However, recently, some high-profile works (Abramson et al., 2024; Wang et al., 2024) have begun to question the need of using equivariant models for certain tasks in light of the ability of high-capacity unconstrained models to achieve similar performance without the intrinsic limitations of architectures with hard-coded equivariance, raising the question of whether equivariance actually matters at scale (Brehmer et al., 2024). These limitations include implementation complexity, scalability issues, excessively constrained hypothesis space (for tasks that involve approximate equivariance due to noisy or imperfect data), or requiring a priori knowledge of the symmetry group.

Besides ignoring the equivariance and using high-capacity fully unconstrained models, the other alternative to ‘hard’ equivariant models is ‘soft’ equivariant models, i.e., unconstrained architectures for which the training protocol is adjusted to encourage equivariance. The most common such approach is data augmentation (Gerken et al., 2022; Quiroga et al., 2020; Xu et al., 2023), but recently, multi-task learning has also been proposed as a way of controlling the extent to which the base model is encouraged to be approximately equivariant (Elhag et al., 2024).

However, these two approaches come with their own limitations. On the one hand, ‘soft’ equivariant models still demand a priori additional knowledge, for instance, about the symmetry group, which often may not be readily available or require domain expertise. On the other hand, fully unconstrained models lack guarantees for learning any relevant symmetry, and they also lack explicit interpretability of their internal representations to allow ascertaining such behavior.

This raises the question of whether it is possible to learn symmetry-aware models but without direct supervision on the symmetry. A recent line of work has explored the problem of automatic symmetry discovery, i.e., explicitly discovering unknown symmetry groups to which a dataset is equivariant solely from the data (Yang et al., 2023; Benton et al., 2020; Yang et al., 2024; Desai et al., 2022). Nevertheless, so far, most of these approaches have been devised as a single step in a model design pipeline, where the symmetry discovery approach is decoupled from the downstream task and hence cannot adjust if only specific symmetries are relevant for the task. Moreover, unsupervised approaches may need any symmetry in the downstream task (e.g., rotation-invariant labels) to be a symmetry in the data (feature) distribution (e.g., the data needs to be uniformly distributed over rotation angles).

In this paper, we propose SEMoLA (Soft-Equivariant Models with Learnable Augmentations), an approach that seeks to overcome these limitations by allowing equivariance to an unknown symmetry to be automatically learned from data while producing predictions through an interpretable end-to-end method. The main contributions of our work are:

- We introduce an end-to-end approach that discovers a priori unknown continuous symmetries using learnable data augmentations, which are used to jointly encode approximate equivariance to the discovered symmetries in an arbitrary underlying unconstrained model.
- Our results show that this approach can improve the performance, sample efficiency, and robustness out-of-distribution of symmetry-agnostic models, and are comparable to those of ‘hard’ and ‘soft’ equivariant models designed with the ground truth symmetry.
- In addition, our approach learns equivariance robustly from both the data and task, while being interpretable by containing an explicit representation of the learned symmetry, and allowing control over how much the equivariance is enforced into the model.

2 RELATED WORK

Equivariant machine learning. Many works have explored the design of equivariant models for groups known a priori, such as permutations in sets (Zaheer et al., 2017), translations (Krizhevsky et al., 2012), roto-translations (Fuchs et al., 2020), or local gauge transformations (Cohen et al., 2019). There are also more general approaches that can be adapted to different specified symmetry groups, such as EMLP (Finzi et al., 2021) and GCNN (Cohen and Welling, 2016). Instead of designing a constrained equivariant model (hard equivariance), other approaches opt for learning the symmetries (soft equivariance) from data through pre-specified augmentations (Hu et al., 2021; Gerken et al., 2022; Wang et al., 2022; Quiroga et al., 2020; Chen et al., 2020; Xu et al., 2023) or multi-task learning (Elhag et al., 2024). Recent works (Abramson et al., 2024; Wang et al., 2024) have shown that large-scale unconstrained architectures can be a valid alternative when having access to enough data. While this last approach is similar to our proposal in that it does not consider a pre-specified symmetry, it fails to provide any insight into what equivariance, if any, the models learn. Another related approach is the soft equivariant model REMUL (Elhag et al., 2024), which frames learning equivariance as a weighted multi-task problem for unconstrained models. However, while that method considers a known symmetry group, our proposal tackles the more complex scenario of discovering the correct symmetry while jointly encoding it into the unconstrained model. Another set of related works correspond to the unsupervised representation learning of latent spaces that disentangle the action of a potentially unknown group on the original input space to facilitate the application of simpler equivariant models that operate on that pretrained latent space (Koyama et al., 2023; Mitchel et al., 2024). These works, however, are not designed to produce interpretable representations of the possibly unknown group or to directly generate approximately equivariant predictions, and, thus, also differ from the problems that our method aims to address.

Symmetry discovery. Recently, multiple methods for automatically discovering symmetries in datasets across different search spaces have been proposed, e.g., Zhou et al. (2021); Karjol et al. (2024) discover discrete finite groups, Augerino (Benton et al., 2020) learns subsets of specified groups, SymmetryGAN (Desai et al., 2022) learns individual symmetry transformations on the dataset distribution, LieGAN (Yang et al., 2023) discovers continuous symmetries, and LaLiGAN (Yang et al., 2024) learns non-linear continuous symmetries. These approaches were introduced as a first step in a model design pipeline, where, after training an initial model to discover the symmetry, this symmetry is used to define an equivariant model that produces predictions. Similarly, other works

108 explore the reversed scenario of defining models that identify symmetries from a previously trained
 109 prediction model (Krippendorf and Syvaeri, 2020; Moskalev et al., 2023; Hu et al., 2024). Other
 110 approaches attempt to discover symmetries through learnable weight-sharing schemes (der Linden
 111 et al., 2024), specifically designed architectures (Dehmamy et al., 2021), or approximate Bayesian
 112 model selection (van der Ouderaa et al., 2024). However, they suffer from additional complexity
 113 and assumptions, limited scalability, and reduced interpretability. For instance, L-conv (Dehmamy
 114 et al., 2021) suffers from high computational complexity due to the computation of exponential maps
 115 in every layer of their hard-constrained architecture, which is also not well-suited for dealing with
 116 approximate and inexact symmetries. Most related to our work are LieGAN (Yang et al., 2023) and
 117 Augerino (Benton et al., 2020), since they also discover continuous symmetries with Lie algebra
 118 bases in an architecture-agnostic manner. However, LieGAN (Yang et al., 2023) does not produce
 119 predictions and discovers symmetries as transformations that preserve the data distribution, which
 120 makes its performance sensitive to the relevant transformations being uniformly distributed in the
 121 data (Hu et al., 2024), while our approach is more robust due to the additional learning signal from the
 122 prediction task. Augerino (Benton et al., 2020) produces predictions but is restricted to discovering
 123 a distribution over coefficients associated with the generators of a given group rather than a fully
 124 unknown group, as we propose, and its training and inference differ from ours, e.g., it does not allow
 125 control over how much equivariance is enforced into the architecture, thus lacking adaptability to
 126 different datasets and scenarios. A modified version, Augerino+, was introduced in (Yang et al.,
 127 2023) to explore its ability to learn unknown symmetry groups, but the experimental results in that
 paper suggest that it is not applicable for this purpose.

128
 129 **Learnable augmentations.** An integral part of our proposal consists of defining a data augmentation
 130 process that, rather than relying on predefined transformations, as is predominantly common in the
 131 literature (Gerken et al., 2022; Quiroga et al., 2020; Xu et al., 2023), has learnable parameters that are
 132 trained jointly with the prediction model. Learnable augmentations have already been explored in the
 133 context, for instance, of anomaly detection for time series (Qiu et al., 2021), or self-supervised GNNs
 134 (Kefato et al., 2021), showing that they are capable of providing additional flexibility and enhanced
 135 performance. Nevertheless, in previous works, the learnable augmentations are usually represented
 136 by black-box models, such as Multi-Layer Perceptrons (MLPs), which hinders the interpretability of
 137 the generated augmentations. In contrast, our approach parameterizes the learnable augmentations as
 138 a Lie algebra basis, as is common in symmetry discovery methods (Yang et al., 2023; 2024; Benton
 139 et al., 2020), thus enabling a straightforward interpretation of the augmentations, which allows for
 greater insight into characteristics of the data that could be of interest to practitioners.

140 3 BACKGROUND

141
 142
 143 **Equivariant models.** We assume that we work with a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ that lives in a
 144 geometric domain with a symmetry group G . Models $f : \mathcal{X} \rightarrow \mathcal{Y}$ from the input domain $\mathcal{X} \subseteq \mathbb{R}^n$
 145 to the output domain $\mathcal{Y} \subseteq \mathbb{R}^m$ are equivariant to the group if the actions of G on \mathcal{X} correspond to
 146 related actions on \mathcal{Y} , through the appropriate group representations:

147 **Definition 3.1.** A function $f : \mathcal{X} \rightarrow \mathcal{Y}$ is equivariant with respect to a group G if $\forall g \in G, (x, y) \in$
 148 \mathcal{D} we have $f(\rho_{\mathcal{X}}(g)x) = \rho_{\mathcal{Y}}(g)f(x)$, where $\rho_{\mathcal{X}} : G \rightarrow \text{GL}(n)$ and $\rho_{\mathcal{Y}} : G \rightarrow \text{GL}(m)$ are
 149 representations of the group G acting on the input \mathcal{X} and output \mathcal{Y} , respectively. Invariance can be
 150 seen as a special case of equivariance, where $\rho_{\mathcal{Y}} = \text{Id}$.

151 **Definition 3.2.** A group representation $\rho : G \rightarrow \text{GL}(n)$ is a group homomorphism that maps every
 152 group element $g \in G$ to a non-singular matrix $\rho(g) \in \mathbb{R}^{n \times n}$ that can be applied to transform the data
 153 from \mathcal{X} . We will sometimes omit $\rho_{\mathcal{X}}$ and $\rho_{\mathcal{Y}}$ when clear and simply write gx or gy instead.

154
 155 **Lie groups.** Our approach focuses on learning continuous symmetries, which are most naturally
 156 modeled by Lie groups, i.e., groups that are also differentiable manifolds, due to their relevance
 157 for many real-world tasks, especially in scientific domains. Some examples include rotation groups
 158 $SO(n)$, or Euclidean groups $E(n)$. Many relevant Lie groups are matrix Lie groups, since they can
 159 be expressed as subgroups of the Lie group $\text{GL}(n, \mathbb{R})$, and their group product is the matrix product.
 160 Each Lie group G is associated with a Lie algebra, which is the tangent space at the identity $\mathfrak{g} = T_{\text{Id}}G$,
 161 and is a vector space of the same dimensionality as the group. A basis of the Lie algebra $L_i \in \mathfrak{g}$ can
 be viewed as the set of infinitesimal generators of the associated Lie group G . To map Lie algebra

elements to elements of the Lie group, we can apply the exponential map $\exp : \mathfrak{g} \rightarrow G$, which, for instance, for matrix groups corresponds to the matrix exponential, i.e., $\exp(A) = \sum_{n=0}^{\infty} \frac{1}{n!} A^n$. If the Lie group G is connected and compact, we can write its elements as $g = \exp(\sum_i w_i L_i)$, with this map being surjective. This parameterization using Lie algebras, which coincides with that of other existing continuous symmetry discovery methods (e.g., LieGAN (Yang et al., 2023) and Augerino (Benton et al., 2020)), therefore has the limitation that it can only represent a single connected component of the associated Lie groups. For non-connected groups, the description of all their group elements would require a more complex parameterization, which, like other existing symmetry discovery methods, we do not consider in this work.

4 METHOD

In this section, we present our approach, SEMoLA (Soft-Equivariant Models with Learnable Augmentations), illustrated in Figure 1. Its main components are the learnable LieAugmenter module that generates augmentations from the original input data using the discovered symmetry, and an arbitrary unconstrained base model that generates the predictions. In the following, we provide a more detailed description of these components as well as the training and inference processes.

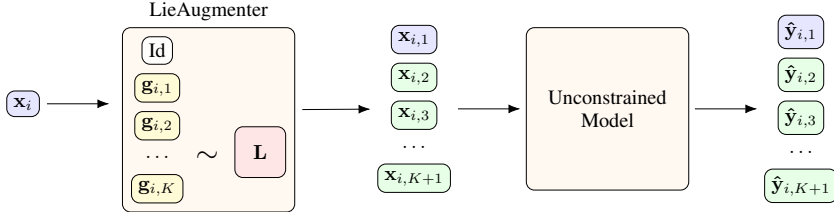


Figure 1: Structure of SEMoLA. The LieAugmenter module learns a Lie algebra basis, from which Lie group elements are sampled to augment the original input data. An arbitrary unconstrained base model then processes both the original and augmented data and outputs the corresponding predictions.

4.1 LIEAUGMENTER

Our learnable augmenter module models continuous symmetries using Lie theory, following an approach similar to some symmetry discovery methods (Benton et al., 2020; Yang et al., 2023), which, while limited to connected Lie groups, is applicable to a wide range of relevant and common symmetry groups (e.g., $SO(n)$, $SE(n)$). Moreover, through alternative parameterization strategies, it can also be applied to other types of symmetries, such as discrete subgroups, as shown in Appendix A.8.

In particular, we reparameterize a distribution over the elements of a matrix Lie group G through its generators, which are given by a Lie algebra basis with cardinality C and dimensionality D , $L = \{L_i \in \mathbb{R}^{D \times D}\}_{i=1}^C$, whose entries we define as parameters that are optimized during training. The value of D is known from the dimensionality of the data, but C needs to be estimated in general. In our experiments, we assume that the value of C for the correct symmetry group is known, but in Appendix A.9 we also show that our approach can be robust to misspecifications.

We can then sample elements from the Lie group by sampling coefficients from a distribution $w \sim P(\gamma)$, and then applying the matrix exponential to the resulting linear combination of elements of the Lie algebra basis $g = \exp\left[\sum_{i=1}^C w_i L_i\right]$ (Falorsi et al., 2019). Unless otherwise stated, in our experiments, we sample the weights from a uniform distribution $P(\gamma) = \mathcal{U}[-\gamma, \gamma]$ to avoid any unwanted bias towards sampling a specific subset of group elements.

The LieAugmenter samples K group elements per input sample, and computes the augmentations as the product of each sampled group element $g \in G$ and the corresponding input sample $x \in \mathcal{X}$. We explore the impact of the chosen number of augmentations on the performance of the approach in Appendix A.9, showing that our method is robust to that choice. We initialize all the parameters of the basis to a constant small value, e.g., 1e-2, to provide an uninformative prior on its potential structure, so that the learning process is not influenced by potentially beneficial or adversarial random

initializations. In addition, we normalize the parameters as $L'_i = \frac{\sqrt{D}}{\|L_i\|} L_i$, with $\|L_i\|$ being the Frobenius norm. This way, the basis maintains a constant norm throughout training for greater stability. The scale factor depends on the dimensionality of the basis, and thus encourages a more consistent representation across groups of different sizes.

4.2 BASE UNCONSTRAINED MODEL

After sampling augmentations of the input data from the LieAugmenter, these are fed into a base model f_θ , which, as we previously mentioned, can be any unconstrained architecture as long as it is compatible with the considered data and task. For instance, in our main experiments, we use a CNN model when working with image data and an MLP for tensor inputs. Nevertheless, there is no a priori restriction on substituting such models for other more complex alternatives, and we explore some of them, e.g., ViTs (Dosovitskiy et al., 2020) and EfficientNets (Tan and Le, 2019), in Appendix A.7. In fact, a model that provides good performance for the task may lead to a faster convergence of the Lie algebra in the LieAugmenter to the correct symmetry, since the loss would be minimized more easily.

4.3 TRAINING

We formulate the training loss as a weighted multi-task objective that balances multiple considerations:

$$\mathcal{L}_{total}(f_\theta, \mathcal{X}, \mathcal{Y}, G, L) = \alpha \mathcal{L}_{obj}(f_\theta, \mathcal{X}, \mathcal{Y}) + \beta \mathcal{L}_{equiv}(f_\theta, \mathcal{X}, \mathcal{Y}, G) + \lambda l_{areg}(\mathcal{X}, G) + \eta l_{bcreg}(L) + \nu l_{bsreg}(L) \quad (1)$$

The two main loss terms are the empirical loss \mathcal{L}_{obj} (e.g., MSE, binary cross-entropy) and the equivariance loss $\mathcal{L}_{equiv}(f_\theta, \mathcal{X}, \mathcal{Y}, G) = \frac{1}{N} \sum_{i=1}^N \frac{1}{K} \sum_{j=1}^K \ell(f_\theta(\rho_{\mathcal{X}}(g_{i,j})(x_i)), \rho_{\mathcal{Y}}(g_{i,j})(y_i))$, with ℓ being a metric, such as the L_1 norm, and $\rho_{\mathcal{Y}}(g_{i,j})$ being simply $g_{i,j}$ for equivariant tasks and the identity for invariant tasks. In addition, we include some regularization terms that focus on encouraging desirable properties of the Lie algebra basis learned by the LieAugmenter module: $l_{areg}(\mathcal{X}, G) = \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^K \text{CosSim}(x_i, g_{i,j}x_i)$ encourages non-trivial augmentations, $l_{bcreg}(L) = \sum_{i=1}^C \sum_{j=i+1}^C \text{CosSim}(L_i, L_j)$ orthogonality of the elements in the basis, and $l_{bsreg}(L) = \sum_{i=1}^C |L_i|$ sparse basis representations. We generally set $\alpha, \beta > \lambda, \eta > \nu > 0$ according to the importance of each term, and we explore ablations of the values for those coefficients in Appendix A.9, which show that our method is very robust to the choice of coefficient values.

The intuition behind why this loss function may allow the LieAugmenter to learn the correct symmetry relies on the fact that its minimization is facilitated when the learned symmetry is compatible with that of the dataset, since otherwise minimizing \mathcal{L}_{equiv} leads to a restriction of the hypothesis space that can negatively affect the ability of the model to minimize \mathcal{L}_{obj} .

4.4 INFERENCE

At inference time, there are two main potential strategies for generating predictions. The first, more robust strategy averages the inversely transformed outputs of the model for the augmented inputs, i.e.,

$$\hat{y}_i = \frac{1}{K+1} \left(\hat{y}_{i,1} + \sum_{j=1}^K \rho_{\mathcal{Y}}(g_{i,j}^{-1}) \hat{y}_{i,j+1} \right), \quad (2)$$

where again $\rho_{\mathcal{Y}}(g_{i,j}^{-1})$ is equal to $g_{i,j}^{-1}$ for equivariant tasks and the identity for invariant ones. This group action is well-defined following the existence of a well-defined action on the ground truth labels and the training objective of our method, which mainly through the equivariance loss term, encourages the model to generate outputs that are consistent with the learned group actions.

The second strategy does not generate any augmentations at inference time and simply processes the original inputs directly with the base model, which might be preferable in scenarios where inference speed is paramount. In our experiments, we mainly apply the averaging strategy in Equation 2, but we also provide some results with the second strategy in Appendix A.4.3, which help confirm the expected trade-off of faster inference speed but slightly worse performance.

270 5 EXPERIMENTS

271
272 In our experiments, we probe how well SEMoLA recovers underlying symmetries, how equivariant
273 the prediction model becomes, and its prediction performance on the target task, both in- and out-of-
274 distribution. We test different underlying symmetries, prediction tasks, and data types. For further
275 details about the experimental setting as well as additional results, please refer to Appendix A. In
276 addition, to facilitate the reproduction of our experiments, we make our implementation available¹.
277

278 **Metrics.** To evaluate the symmetry discovery results of the different approaches, we mainly rely
279 on a qualitative comparison of visualizations of the learned Lie algebra bases, but we also report
280 approximate quantitative measurements in the form of primarily the absolute cosine similarity and
281 also the Mean Absolute Error (MAE) with the ground truth. These quantitative metrics are not always
282 fully reliable because they ignore alternative representations of the Lie algebra basis of the correct
283 symmetry, but they still provide some insight into the robustness of the approaches. Nevertheless,
284 we complement them by also considering the effect that applying augmentations from the learned
285 symmetries has on the base model’s performance compared to using the ground truth symmetries,
286 which we believe provides a strong indication of the correctness of the discovered symmetries.

287 To quantify the equivariance learned by the models, similarly to Elhag et al. (2024), we apply the
288 following equivariant error metric:

$$289 E(f_\theta, G) = \frac{1}{N} \sum_{i=1}^N \left\| \frac{1}{K} \sum_{j=1}^K \rho_{\mathcal{Y}}(g_j)(f_\theta(x_i)) - \frac{1}{K} \sum_{j=1}^K f_\theta(\rho_{\mathcal{X}}(g_j)(x_i)) \right\| \quad (3)$$

292 where $\|\cdot\|$ is defined to be the L_1 norm, and we consider $K = 10$.
293

294 **Baselines.** When analyzing the performance of our method, it is crucial to compare it against
295 relevant baselines from the literature. However, to the best of our knowledge, there is no existing
296 alternative that successfully performs interpretable discovery of continuous symmetries and equiv-
297 ariance learning while producing predictions with an unconstrained model, as we do. Thus, when
298 defining the baselines, we can only consider approaches designed to address a subset of those tasks.
299

300 For the symmetry discovery task, we provide a comparison with LieGAN and Augerino+ (Yang
301 et al., 2023). Then, for equivariance and prediction performance, we consider two ‘hard’ equivariant
302 models that are adaptable to different symmetry groups, GCNN (Cohen and Welling, 2016) and
303 EMLP (Finzi et al., 2021). We also provide the performance of the unconstrained model that we use
304 for SEMoLA without the augmenter and also with a non-learnable augmenter that uses the ground
305 truth Lie algebra basis, to establish approximate lower and upper bounds on the performance and
306 equivariance achievable by our proposal. Lastly, we also report the performance of incorporating
307 the Lie algebra basis discovered by LieGAN into both EMLP and the non-learnable augmenter,
308 seeking to demonstrate the performance that the two-step process of discovering and then encoding a
309 symmetry can accomplish.

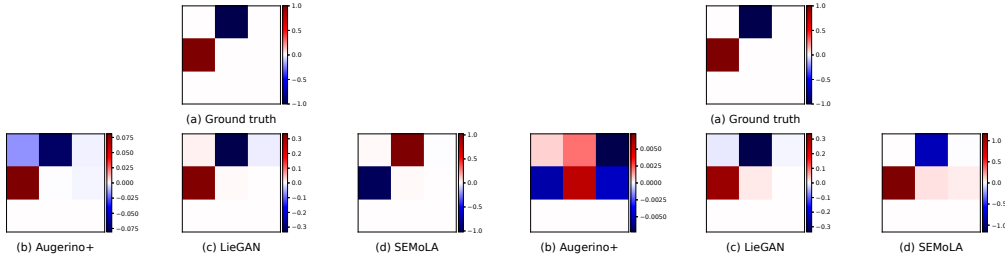
310 5.1 ROTATEDMNIST

311 Firstly, we consider a version of the MNIST dataset (Deng, 2012) where we randomly rotate each
312 image, thus establishing the rotation group $SO(2)$ as the underlying symmetry we want to discover.
313 In particular, we define two scenarios: in-distribution (in which the images of the train and test sets
314 are randomly rotated by an angle in $[0, 360)$ degrees), and out-of-distribution (in which the images
315 of the training set are randomly rotated by an angle in $[0, 90) \cup [0, -90)$ degrees and for the test
316 set the rotation is by an angle in $[90, 180) \cup [-90, -180)$ degrees). This out-of-distribution setting
317 mainly affects the symmetry discovery methods in that they have to learn the correct symmetry from
318 a restricted and less diversely augmented training set, and the predictive models in that they need to
319 be robust to differently distributed data between training and testing.

320 For the symmetry discovery methods, including SEMoLA’s LieAugmenter module, we set the
321 cardinality of the Lie algebra basis to $C = 1$ and search over the six-dimensional space of 2D
322 affine transformations, which are defined to act on the pixel coordinates of the images. Even though
323

¹<https://anonymous.4open.science/r/SEMoLA/>

324 this assumption introduces some prior knowledge into this particular symmetry discovery task, we
 325 believe that it still leads to meaningful results since it corresponds to common practical scenarios.
 326 As we can observe in Figure 2, SEMoLA and LieGAN can identify the correct symmetry for this
 327 dataset in both the in-distribution and out-of-distribution scenarios up to a sign (which does not
 328 impact the correctness since the sampling distribution $P(\gamma)$ is symmetric around 0), while Augerino+
 329 performs worse in the in-distribution case, and fails to learn the correct Lie algebra basis in the
 330 out-of-distribution scenario. Moreover, from Table 1, we can see that while SEMoLA robustly
 331 discovers the correct symmetry in both the in-distribution and out-of-distribution cases, LieGAN
 332 is only robust in-distribution, i.e., it is sensitive to the non-uniform distribution of rotations in the
 333 out-of-distribution data, which is likely due to its objective for symmetry discovery that focuses on
 334 transformations that preserve the data distribution.



344 Figure 2: Comparison of the ground truth Lie algebra basis with those learned by LieGAN, Augerino+,
 345 and our proposed SEMoLA approach for the in-distribution version of the RotatedMNIST dataset
 346 (left) and the restricted out-of-distribution version (right).
 347

349 Table 1: Cosine and MAE similarities of the ground truth Lie algebra basis and the ones learned
 350 by each of the considered symmetry learning models in the in-distribution and out-of-distribution
 351 versions of the RotatedMNIST dataset. For each metric, we report the mean and standard deviation
 352 across three runs using different random seeds.

	In distribution		Out of distribution	
Model	Cosine	MAE	Cosine	MAE
Augerino+	0.7883±0.1407	0.2169±0.0068	0.3418±0.2208	0.2246±0.0004
LieGAN	0.9998±0.0002	0.1379±0.0104	0.7490±0.3540	0.1705±0.0379
SEMoLA	0.9997±0.0002	0.3077±0.1989	0.9932±0.0063	0.1940±0.1956

360 In terms of classification performance, Table 2 shows that SEMoLA provides the overall second-best
 361 performance, only behind (but close to) the performance of the CNN model that considers ground
 362 truth augmentations. The CNN model with augmentations from the Lie algebra basis discovered by
 363 LieGAN also shows a similarly high performance, while the other baselines, i.e., Augerino+, CNN,
 364 and GCNN, all provide considerably lower accuracy in both scenarios. The performance of GCNN is
 365 perhaps unexpectedly low for a hard equivariant model, but the reason probably lies in its design for
 366 invariance to the $p4$ group, which does not exactly match the continuous symmetry of this dataset.

367 We provide further results for this dataset in Appendix A.4, such as analyzing the sample efficiency of
 368 the prediction approaches, from which we can see that our proposal mostly matches the performance
 369 of the base model with ground truth augmentations even with decreasing sizes of the training set.
 370

371 5.2 N-BODY DYNAMICS

372
 373 As the next experiment, we consider a simulated N-body trajectory dataset from Greydanus et al.
 374 (2019), generated as in Yang et al. (2023). It consists of two bodies with identical masses rotating
 375 in nearly circular orbits, and the objective is to predict future coordinates based on past observa-
 376 tions, which is a rotationally equivariant problem. In general, the input and output feature vectors
 377 for each timestep have $4N$ entries, corresponding to the positions and momenta of the bodies:
 $[q_{1x}, q_{1y}, p_{1x}, p_{1y}, \dots, q_{Nx}, q_{Ny}, p_{Nx}, p_{Ny}]$. Again, we define two scenarios: in-distribution (i.e.,

Table 2: Results for the classification performance and equivariant error of the considered models for in-distribution and out-of-distribution versions of the RotatedMNIST dataset. For each metric, we report the mean and standard deviation across three runs using different random seeds. The accuracies are expressed as percentages.

Model	In distribution		Out of distribution	
	Accuracy	Equiv. Error	Accuracy	Equiv. Error
Augerino+	96.85±0.30	5.09±0.70	57.68±1.45	3.97±0.51
CNN	96.17±0.34	5.52±0.23	55.24±1.14	7.19±0.54
CNN + GT aug	99.06±0.00	4.01±0.29	99.09±0.08	3.87±0.35
CNN + LieGAN aug	98.69±0.16	5.04±0.19	91.59±1.07	5.82±0.47
GCNN	96.78±0.36	5.31±1.16	55.60±1.45	5.78±1.23
SEMoLA	99.08±0.06	3.10±0.32	90.82±6.49	3.26±0.62

the default dataset), and out-of-distribution (in which the training set is modified to ensure that the first body’s position is always in the top-left or bottom-right quadrants, and for the test set it is always in the top-right or bottom-left quadrants). Here, we focus our analysis on the task of predicting one output timestep from one input timestep, but we provide further results in Appendix A.5, including for the more challenging scenario of considering three input and output timesteps, which show that our approach performs similarly well in that setting.

As we can see in Figure 3, both LieGAN and Augerino+ fail to learn the correct symmetry. Conversely, SEMoLA discovers an alternative representation of the Lie algebra basis of the correct symmetry under the constraints of the considered dataset (see Appendix A.5.5). LieGAN is capable of identifying the correct symmetry but only in the in-distribution case and when complemented with auxiliary knowledge in the form of different masks that restrict the interaction between different features of the input (see Appendix A.5), while Augerino+ still fails.

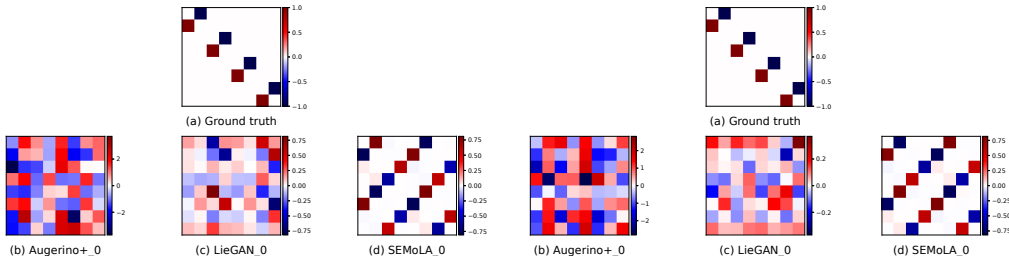


Figure 3: Comparison of the ground truth Lie algebra basis with those learned by LieGAN, Augerino+, and our proposed SEMoLA approach for the in-distribution version of the 2-body dataset (left) and the restricted out-of-distribution version (right) with one input and output timestep.

From the results in Table 3, we can see that SEMoLA provides the highest performance behind the hard-equivariant baselines that consider pre-specified symmetries, while outperforming Augerino+ and the base MLP (even when it considers augmentations from pre-specified symmetries) in terms of both task performance and equivariant error across in-distribution and out-of-distribution scenarios. We should note that the fact that SEMoLA performs slightly better than the MLP with ground truth augmentations is probably due to the alternative representation of the symmetry that it learns, which encourages more diverse rotations, as shown in Appendix A.5.5.

5.3 QM9

We also explore the performance of SEMoLA on the more complex task of molecular property prediction using the QM9 dataset (Blum and Raymond, 2009; Rupp et al., 2012), which consists of small inorganic molecules described by their 3D coordinates and atomic charges. We focus

Table 3: Results for the performance and equivariant errors of the considered models for in-distribution and out-of-distribution versions of the 2-body dataset with one input and output timestep. For each metric, we report the mean and standard deviation across three runs using different random seeds.

Model	In distribution		Out of distribution	
	MSE	Equiv. Error	MSE	Equiv. Error
Augerino+	$5.28e+00 \pm 4.23e-01$	$1.90e-01 \pm 1.92e-03$	$3.26e+00 \pm 7.30e-01$	$1.94e-01 \pm 4.15e-03$
MLP	$5.16e-03 \pm 2.79e-03$	$2.48e-02 \pm 7.37e-03$	$1.86e-02 \pm 3.74e-03$	$3.33e-02 \pm 2.56e-03$
MLP + GT aug	$2.13e-05 \pm 1.56e-05$	$3.03e-03 \pm 1.03e-03$	$3.03e-05 \pm 8.81e-06$	$3.70e-03 \pm 1.14e-03$
MLP + LieGAN aug	$3.23e-03 \pm 6.25e-04$	$1.91e-02 \pm 1.47e-03$	$1.15e-02 \pm 3.00e-03$	$2.84e-02 \pm 3.04e-03$
EMLP + GT	$1.74e-06 \pm 1.50e-06$	$5.32e-05 \pm 1.16e-05$	$4.64e-06 \pm 4.86e-06$	$5.27e-05 \pm 1.10e-05$
EMLP + LieGAN	$6.36e-06 \pm 4.56e-06$	$5.31e-05 \pm 1.13e-05$	$3.81e-06 \pm 2.62e-06$	$5.31e-05 \pm 1.09e-05$
SEMOLA	$1.39e-05 \pm 6.08e-07$	$2.78e-03 \pm 4.88e-04$	$2.38e-05 \pm 9.04e-06$	$2.59e-03 \pm 9.89e-05$

on the tasks of predicting the highest occupied molecular orbital energy (HOMO) and the lowest unoccupied molecular orbital energy (LUMO), which are invariant under transformations of the atomic coordinates given by the SE(3) group. Thus, they provide an ideal scenario for testing the capability of our approach for learning more complex symmetries, consisting of multiple elements in their Lie algebra basis. We consider the LieConv model (Finzi et al., 2020) with no equivariance (LieConv-Trivial) as the base model for our SEMoLA approach, and compare its performance with that reported for different augmentation strategies, including using the ground truth and the symmetries discovered by LieGAN and Augerino (Benton et al., 2020).

As we can see in Figure 4, SEMoLA learns the correct Lie algebra basis of the SE(3) group up to permutation (which is correct due to the permutation-invariant nature of our strategy for sampling group elements) for both prediction tasks, and from the results in Table 4, we see that its performance is comparable to that of augmentations with the ground truth symmetry and that it outperforms using symmetries learned by LieGAN and Augerino.

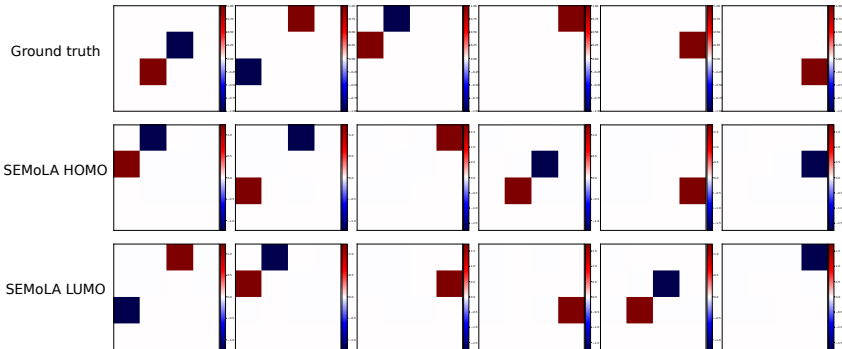


Figure 4: Comparison of the ground truth Lie algebra basis and those learned by SEMoLA for the HOMO and LUMO targets of the QM9 dataset.

6 CONCLUSION

In this paper, we present SEMoLA, an approach for defining learnable equivariant models without prior knowledge of the underlying symmetry, while providing interpretability, high-performant predictions, and low equivariance errors, as shown across multiple experiments. In particular, our results show that our method is able to match or surpass the performance of models specifically designed for either symmetry discovery or equivariant predictions, while jointly tackling both tasks.

While hard equivariant models still have many advantages (e.g., exact equivariance guarantees, higher sample and parameter efficiency, etc.) in cases where the symmetry is known a priori and exact, their adoption is limited by the complexity of their design and implementation. In this context, our

Table 4: Results for the performance and equivariant errors in meVs of the different considered models for the QM9 dataset. The results for the models marked with [†] are from Benton et al. (2020) and those marked with [‡] come from Yang et al. (2023).

Model	HOMO		LUMO	
	MAE	Equiv. Error	MAE	Equiv. Error
No Symmetry [†]	52.7	-	43.5	-
SE(3) [†]	36.5	-	29.8	-
Augerino [†]	38.3	-	33.7	-
LieGAN [‡]	43.5	-	36.4	-
SEMoLA	37.09	0.0033	30.67	0.0015

work can be seen as an initial attempt at streamlining the design, flexibility, interpretability, and applicability of equivariant models, which we hope can lead to a more widespread adoption of these models, especially when the underlying symmetries are inexact or harder to ascertain, as is quite common for real-world tasks in scientific domains, such as physics or chemistry.

Nevertheless, there are still many potential research directions based on our approach that could be explored in future work. For instance, modeling learnable non-connected Lie groups and non-linear symmetries, or applying our method to tasks without clearly known symmetries.

REPRODUCIBILITY STATEMENT

We believe that our results being reproducible is of great importance to the correctness and usefulness of our work, so we make an effort to provide as many details as possible to ensure their reproducibility. In that sense, our description of the proposed method in Section 4 and of the experiments and results in Section 5, together with the additional details presented throughout Appendix A, covers the most relevant information for the reproduction of the results, including, for instance, the specific hyperparameter values considered in each of the experiments. In addition, we provide access to our code (<https://anonymous.4open.science/r/SEMoLA/>), which includes the configuration files used in the execution of all the experiments, as well as instructions on how to access the different datasets, and we also plan to publicly release it upon acceptance of the paper.

REFERENCES

- Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O’Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Židek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w.
- Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarakar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhersch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster

- 540 Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation.
541 In *29th ACM International Conference on Architectural Support for Programming Languages and*
542 *Operating Systems, Volume 2 (ASPLOS '24)*. ACM, April 2024. doi: 10.1145/3620665.3640366.
543 URL <https://pytorch.org/assets/pytorch2-2.pdf>.
- 544 Joseph Aylett-Bullock, Simon Badger, and Ryan Moodie. Optimising simulations for diphoton
545 production at hadron colliders using amplitude neural networks. *Journal of High Energy Physics*,
546 2021(8):1–30, 2021.
- 548 Simon Badger and Joseph Bullock. Using neural networks for efficient evaluation of high multiplicity
549 scattering amplitudes. *Journal of High Energy Physics*, 2020(6):1–26, 2020.
- 550 Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew G Wilson. Learning Invariances in Neural
551 Networks from Training Data. In *Advances in Neural Information Processing Systems*, volume 33,
552 pages 17605–17616. Curran Associates, Inc., 2020.
- 554 Lorenz C Blum and Jean-Louis Reymond. 970 million druglike small molecules for virtual screening
555 in the chemical universe database GDB-13. *Journal of the American Chemical Society*, 131(25):
556 8732–8733, 2009.
- 557 Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J Bekkers, and Max Welling. Ge-
558 ometric and physical quantities improve e(3) equivariant message passing. *arXiv preprint*
559 *arXiv:2110.02905*, 2021.
- 561 Johann Brehmer, Sönke Behrends, Pim de Haan, and Taco Cohen. Does Equivariance Matter at
562 Scale?, October 2024.
- 563 Srinath Bulusu, Matteo Favoni, Andreas Ipp, David I Müller, and Daniel Schuh. Generalization
564 capabilities of translationally equivariant neural networks. *Physical Review D*, 104(7):074504,
565 2021.
- 567 Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation.
568 *Journal of Machine Learning Research*, 21(245):1–71, 2020.
- 569 Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional
570 networks and the icosahedral CNN. In *International conference on Machine learning*, pages
571 1321–1330. PMLR, 2019.
- 572 Taco S. Cohen and Max Welling. Group Equivariant Convolutional Networks, June 2016.
- 574 Nima Dehmamy, Robin Walters, Yanchen Liu, Dashun Wang, and Rose Yu. Automatic Symmetry
575 Discovery with Lie Algebra Convolutional Network. In *Advances in Neural Information Processing*
576 *Systems*, volume 34, pages 2503–2515. Curran Associates, Inc., 2021.
- 577 Li Deng. The mnist database of handwritten digit images for machine learning research [best of the
578 web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.
- 580 Putri A. Van der Linden, Alejandro García Castellanos, Sharvaree Vadgama, Thijs P. Kuipers, and
581 Erik J. Bekkers. Learning Symmetries via Weight-Sharing with Doubly Stochastic Tensors. In
582 *ICML 2024 Workshop on Geometry-grounded Representation Learning and Generative Modeling*,
583 June 2024.
- 584 Krish Desai, Benjamin Nachman, and Jesse Thaler. SymmetryGAN: Symmetry Discovery with
585 Deep Learning. *Physical Review D*, 105(9):096031, May 2022. ISSN 2470-0010, 2470-0029. doi:
586 10.1103/PhysRevD.105.096031.
- 588 Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas
589 Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An
590 image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint*
591 *arXiv:2010.11929*, 2020.
- 592 Ahmed A. Elhag, T. Konstantin Rusch, Francesco Di Giovanni, and Michael Bronstein. Relaxed
593 Equivariance via Multitask Learning, October 2024.

- 594 Luca Falorsi, Pim de Haan, Tim R Davidson, and Patrick Forré. Reparameterizing distributions on
595 lie groups. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages
596 3244–3253. PMLR, 2019.
- 597
- 598 Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional
599 neural networks for equivariance to lie groups on arbitrary continuous data. In *International
600 Conference on Machine Learning*, pages 3165–3176. PMLR, 2020.
- 601
- 602 Marc Finzi, Max Welling, and Andrew Gordon Wilson. A practical method for constructing equivari-
603 ant multilayer perceptrons for arbitrary matrix groups. In *International conference on machine
604 learning*, pages 3318–3328. PMLR, 2021.
- 605
- 606 Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-
607 translation equivariant attention networks. *Advances in Neural Information Processing Systems*,
33:1970–1981, 2020.
- 608
- 609 Jan Gerken, Oscar Carlsson, Hampus Linander, Fredrik Ohlsson, Christoffer Petersson, and Daniel
610 Persson. Equivariance versus augmentation for spherical images. In *International Conference on
611 Machine Learning*, pages 7404–7421. PMLR, 2022.
- 612
- 613 Jan E. Gerken and Pan Kessel. Emergent Equivariance in Deep Ensembles, June 2024.
- 614
- 615 Samuel Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian neural networks. *Advances in
616 neural information processing systems*, 32, 2019.
- 617
- 618 Jiaqi Han, Wenbing Huang, Tingyang Xu, and Yu Rong. Equivariant graph hierarchy-based neural
619 networks. *Advances in Neural Information Processing Systems*, 35:9176–9187, 2022.
- 620
- 621 Lexiang Hu, Yikang Li, and Zhouchen Lin. Symmetry Discovery for Different Data Types, October
622 2024.
- 623
- 624 Weihua Hu, Muhammed Shuaibi, Abhishek Das, Siddharth Goyal, Anuroop Sriram, Jure Leskovec,
625 Devi Parikh, and C Lawrence Zitnick. Forcenet: A graph neural network for large-scale quantum
626 calculations. *arXiv preprint arXiv:2103.01436*, 2021.
- 627
- 628 John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger,
629 Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate
630 protein structure prediction with AlphaFold. *nature*, 596(7873):583–589, 2021.
- 631
- 632 Pavan Karjol, Rohan Kashyap, Aditya Gopalan, and A. P. Prathosh. A Unified Framework for
633 Discovering Discrete Symmetries. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors,
634 *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume
635 238 of *Proceedings of Machine Learning Research*, pages 793–801. PMLR, 02–04 May 2024.
636 URL <https://proceedings.mlr.press/v238/karjol24a.html>.
- 637
- 638 Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human col-
639 orectal cancer and healthy tissue, April 2018. URL [https://doi.org/10.5281/zenodo.
1214456](https://doi.org/10.5281/zenodo.1214456).
- 640
- 641 Zekarias T Kefato, Sarunas Girdzijauskas, and Hannes Stärk. Jointly learnable data augmentations
642 for self-supervised gnns. *arXiv preprint arXiv:2108.10420*, 2021.
- 643
- 644 Masanori Koyama, Kenji Fukumizu, Kohei Hayashi, and Takeru Miyato. Neural fourier transform: A
645 general approach to equivariant representation learning. *arXiv preprint arXiv:2305.18484*, 2023.
- 646
- 647 Sven Krippendorf and Marc Syvaeri. Detecting Symmetries with Neural Networks, March 2020.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolu-
tional neural networks. *Advances in neural information processing systems*, 25, 2012.
- Thomas Mitchel, Michael J Taylor, and Vincent Sitzmann. Neural isometries: Taming transformations
for equivariant ml. *Advances in Neural Information Processing Systems*, 37:7311–7338, 2024.

- 648 Artem Moskalev, Anna Sepliarskaia, Ivan Sosnovik, and Arnold Smeulders. LieGG: Studying
649 Learned Lie Group Generators, January 2023.
- 650
- 651 Chen Qiu, Timo Pfommer, Marius Kloft, Stephan Mandt, and Maja Rudolph. Neural transformation
652 learning for deep anomaly detection beyond images. In *International conference on machine*
653 *learning*, pages 8703–8714. PMLR, 2021.
- 654 Facundo Quiroga, Franco Ronchetti, Laura Lanzarini, and Aurelio F Bariviera. Revisiting data
655 augmentation for rotational invariance in convolutional neural networks. In *Modelling and Simu-*
656 *lation in Management Sciences: Proceedings of the International Conference on Modelling and*
657 *Simulation in Management Sciences (MS-18)*, pages 127–141. Springer, 2020.
- 658 Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole Von Lilienfeld. Fast
659 and accurate modeling of molecular atomization energies with machine learning. *Physical review*
660 *letters*, 108(5):058301, 2012.
- 661
- 662 Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks.
663 In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.
- 664 Jonas Spinner, Victor Bresó, Pim De Haan, Tilman Plehn, Jesse Thaler, and Johann Brehmer.
665 Lorentz-equivariant geometric algebra transformers for high-energy physics. *Advances in neural*
666 *information processing systems*, 37:22178–22205, 2024.
- 667
- 668 Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks.
669 In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
- 670 Tycho F. A. van der Ouderaa, Mark van der Wilk, and Pim de Haan. Noether’s Razor: Learning
671 Conserved Quantities, October 2024.
- 672
- 673 Rui Wang, Robin Walters, and Rose Yu. Data augmentation vs. equivariant networks: A theory of
674 generalization on dynamics forecasting. *arXiv preprint arXiv:2206.09450*, 2022.
- 675 Yuyang Wang, Ahmed A. A. Elhag, Navdeep Jaitly, Joshua M. Susskind, and Miguel Ángel Bautista.
676 Swallowing the Bitter Pill: Simplified Scalable Conformer Generation. In *Proceedings of the 41st*
677 *International Conference on Machine Learning*, pages 50400–50418. PMLR, July 2024.
- 678
- 679 Mingle Xu, Sook Yoon, Alvaro Fuentes, and Dong Sun Park. A comprehensive survey of image
680 augmentation techniques for deep learning. *Pattern Recognition*, 137:109347, 2023.
- 681 Minkai Xu, Jiaqi Han, Aaron Lou, Jean Kossaifi, Arvind Ramanathan, Kamyar Azizzadenesheli,
682 Jure Leskovec, Stefano Ermon, and Anima Anandkumar. Equivariant graph neural operator for
683 modeling 3d dynamics. *arXiv preprint arXiv:2401.11037*, 2024.
- 684
- 685 Jianke Yang, Robin Walters, Nima Dehmamy, and Rose Yu. Generative Adversarial Symmetry
686 Discovery. In *Proceedings of the 40th International Conference on Machine Learning*, pages
687 39488–39508. PMLR, July 2023.
- 688 Jianke Yang, Nima Dehmamy, Robin Walters, and Rose Yu. Latent Space Symmetry Discovery,
689 August 2024.
- 690
- 691 Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and
692 Alexander J Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, volume 30.
693 Curran Associates, Inc., 2017.
- 694 Allan Zhou, Tom Knowles, and Chelsea Finn. Meta-learning symmetries by reparameterization,
695 2021. URL <https://arxiv.org/abs/2007.02933>.
- 696
- 697
- 698
- 699
- 700
- 701

A EXPANDED EXPERIMENTAL DETAILS

A.1 TECHNICAL INFORMATION

All experiments reported in this paper were executed on a single NVIDIA A100 GPU (except for the QM9 dataset, for which we used an H100) using Pytorch (Ansel et al., 2024) and CUDA version 12.2. We used the Adam optimizer with a learning rate of 0.001, keeping the rest of its default hyperparameter values. As a note, we should mention that even though we report the train and test times for each of the experiments, we notice that those results are easily influenced by the utilization of the GPU at the time of their execution, so some variability can be expected across experiments.

A.2 HYPERPARAMETER CONFIGURATIONS

In order to facilitate a fair comparison of our approach with the considered baselines, we run them using the same exact hyperparameter configurations provided in their papers or in available implementations. Moreover, we make sure to run all the models with the number of epochs and batch size considered in those configurations, instead of defining those values based on the ones that lead to the best performance from our models. Therefore, it is possible that the results of SEMoLA could be slightly improved with a more detailed tuning and selection of such values. Furthermore, we do not carry out an extensive fine-grained hyperparameter search over those hyperparameters unique to our approach, due to the required computational and temporal resources that would entail. Thus, it is possible that this would be another avenue to further improve the performance of our proposal. Lastly, we should note that the hyperparameter values for the base models and base models with fixed augmentations exactly coincide with those of SEMoLA in all of the experiments.

A.3 ADDITIONAL METRICS

In order to provide a more detailed analysis of the equivariance error incurred by the models, we also report their results in all of the considered experiments for another metric from Elhag et al. (2024):

$$E'(f_\theta, G) = \frac{1}{N} \sum_{i=1}^N \frac{1}{K} \sum_{j=1}^K \|f_\theta(\rho_X(g_j)(x_i)) - \rho_Y(g_j)(f_\theta(x_i))\|, \quad (4)$$

where, again, $\|\cdot\|$ is defined to be the L_1 norm, and we consider $K = 10$. We should note that we find this metric to behave very similarly to the one defined in Equation 3, but we include it nonetheless for completeness.

In addition, we also report the total train and test times that correspond to the different prediction models to complement their associated performance metrics. Importantly, those time measurements do not include the additional time required by decoupled symmetry discovery approaches, e.g., LieGAN, for which the prediction models rely on augmentations derived from previously discovered symmetries. That additional runtime for those two-step equivariant prediction approaches can pose a significant limitation in practice, and it is one of the motivations behind our proposal of a single end-to-end method.

A.4 ROTATEDMNIST

A.4.1 EXPERIMENTAL DETAILS

The base model used in SEMoLA, as well as the other baselines that consider it, is defined to be a very simple CNN, with four convolutional filters and two interleaved max-pooling layers. Following Cohen and Welling (2016), the GCNN model is defined as invariant to the $p4$ group, which consists of all possible compositions of translations and rotations by 90 degrees about any center of rotation in a square grid. Its underlying implementation is defined analogously to our base CNN model in terms of number and type of layers, in order to facilitate the comparison. Also, we should note that while this model encodes the invariance into the architecture in a ‘hard’ manner, its performance may be limited by the fact that the discrete group to which it is invariant is not an exact match to the underlying continuous group of the dataset, so while the equivariant errors should be low, its actual classification performance might be somewhat restricted.

In this experiment, we applied the following hyperparameter values to our model: $\alpha = 1.0$, $\beta = 7.0$, $\lambda = 0.1$, $\eta = 0.0$, $\nu = 0.01$, $\gamma = 3.0$ for the uniform distribution in the LieAugmenter, and number of augmentations $K = 10$. We should note that η has no effect on this particular scenario, because the cardinality of the basis is $C = 1$. In addition, we maintain the default train-test splits of MNIST (Deng, 2012) but also create a validation set with 10% of the training set samples, thus having 54,000 training samples, 6,000 validation samples, and 10,000 test samples. We train all models for 15 epochs with a batch size of 64 for three runs with different random seeds and report the mean and standard deviation of the performances.

A.4.2 DEFAULT INFERENCE STRATEGY

We complement the performance results from the paper (Table 2) with the corresponding total train and test times of the prediction models, together with the second equivariant error (Equation 4) in Table 5. From them, we can see that our model takes the longest time to train, which can mainly be explained by its additional complexity in terms of more trainable parameters or loss terms than the alternatives, as well as for our currently not excessively optimized implementation. In addition, the second equivariant error behaves similarly to the one shown in the paper.

Table 5: Results for the train and test times (MM:SS) and equivariant error from Equation 4 of the different considered models for in-distribution and out-of-distribution versions of the RotatedMNIST dataset. We report the mean of all the metrics and the standard deviation of the equivariant error across three runs using different random seeds.

Model	In distribution			Out of distribution		
	Train time	Test time	Equiv. Error'	Train time	Test time	Equiv. Error'
Augerino+	04 : 30	00 : 03	6.35±0.98	04 : 47	00 : 03	5.26±0.72
CNN	01 : 52	00 : 00	6.69±0.24	01 : 50	00 : 00	8.74±0.61
CNN + GT aug	06 : 55	00 : 02	4.86±0.36	06 : 57	00 : 02	4.75±0.38
CNN + LieGAN aug	06 : 15	00 : 01	6.12±0.25	05 : 03	00 : 01	7.17±0.50
GCNN	03 : 26	00 : 00	6.51±1.41	03 : 56	00 : 00	7.16±1.52
SEMOLA	07 : 47	00 : 01	3.82±0.38	07 : 04	00 : 01	4.12±0.75

A.4.3 ALTERNATIVE INFERENCE STRATEGY

Furthermore, we report the performance of the alternative inference strategy of producing predictions only for the original samples without any augmentations in Tables 6 and 7, from which we can see that, as expected, the performance slightly decreases with respect to the previous inference strategy, but the total test time is reduced.

A.4.4 SAMPLE EFFICIENCY

Next, we explore the sample efficiency of SEMOLA when compared to the other considered prediction approaches by measuring their test performance when trained with decreasing fractions of samples in the training set. As we can see in Figure 5, SEMOLA matches the performance of the CNN with ground truth augmentations up to training with 50% of the data, after which its performance decreases but remains higher than that of the other alternatives.

A.4.5 EMPIRICAL CONVERGENCE

While the difficulty of analyzing the joint minimization problem corresponding to SEMOLA puts an exact theoretical analysis of its convergence beyond the scope of this paper, we carry out an empirical exploration in order to provide greater insight and evidence of the applicability of our proposed approach in practice.

In that sense, we report the train and validation loss curves as well as the corresponding separate evolution of the main loss terms of the train loss (i.e., objective/empirical loss and equivariance loss) during training in Figure 6. As we can observe, all losses quickly decrease in the first five epochs and

Table 6: Results for the performance and equivariant error of the different considered models for in-distribution and out-of-distribution versions of the RotatedMNIST, where inference for SEMoLA and CNN + aug models is computed using the second strategy described in the paper, i.e., not computing any augmentations and simply producing the predictions for the original inputs. For each metric, we report the mean and standard deviation across three runs with different random seeds. The accuracies are expressed as percentages.

Model	In distribution		Out of distribution	
	Accuracy	Equiv. Error	Accuracy	Equiv. Error
Augerino+	96.85±0.30	5.09±0.70	57.68±1.45	3.97±0.51
CNN	96.17±0.34	5.52±0.23	55.24±1.14	7.19±0.54
CNN + GT aug	98.45±0.11	3.91±0.23	98.35±0.11	4.03±0.17
CNN + LieGAN aug	97.84±0.13	4.92±0.74	84.97±0.43	6.80±0.85
GCNN	96.78±0.36	5.31±1.16	55.60±1.45	5.78±1.23
SEMoLA	98.55±0.08	3.72±0.15	93.40±2.46	3.40±0.47

Table 7: Results for the train and test times (MM:SS) and equivariant error from Equation 4 of the different considered models for in-distribution and out-of-distribution versions of the RotatedMNIST dataset, where inference for SEMoLA and CNN + augmentations models is computed using the second strategy described in the paper, i.e., not computing any augmentations and simply producing the predictions for the original inputs. We report the mean of all the metrics and the standard deviation of the equivariant error across three runs using different random seeds.

Model	In distribution			Out of distribution		
	Train time	Test time	Equiv. Error'	Train time	Test time	Equiv. Error'
Augerino+	04 : 30	00 : 03	6.35±0.98	04 : 47	00 : 03	5.26±0.72
CNN	01 : 52	00 : 00	6.69±0.24	01 : 50	00 : 00	8.74±0.61
CNN + GT aug	03 : 03	00 : 00	4.72±0.27	06 : 57	00 : 00	4.96±0.13
CNN + LieGAN aug	09 : 45	00 : 00	5.95±0.87	06 : 18	00 : 00	8.35±0.97
GCNN	03 : 26	00 : 00	6.51±1.41	03 : 56	00 : 00	7.16±1.52
SEMoLA	10 : 44	00 : 00	4.54±0.17	10 : 48	00 : 00	4.29±0.56

then start converging around epoch ten. These observations are also similarly reflected in other runs for this same dataset as well as for other more complex datasets (see, e.g., Appendix A.6).

A.4.6 EXAMPLE SAMPLED AUGMENTATIONS

Finally, we plot some examples of the augmentations generated by SEMoLA for different digits in Figure 7, which serves to verify that it is indeed learning to generate augmentations that correspond to rotations of the input image and that the rotations are varied across the augmented samples.

A.5 N-BODY DYNAMICS

A.5.1 EXPERIMENTAL DETAILS

In this experiment, we used a simple MLP, consisting of four layers and ReLU activations, as the base model for SEMoLA and the other baselines that require it. We consider EMLP Finzi et al. (2021) as the ‘hard’ equivariant model for this problem, which we define based either on the ground truth Lie algebra basis, or on the bases learned by LieGAN.

We applied the following hyperparameter values to our model: $\alpha = 1.0$, $\beta = 10.0$, $\lambda = 1.0$, $\eta = 0.0$, $\nu = 0.001$, $\gamma = 2.0$ for the uniform distribution in the LieAugmenter, and number of augmentations $K = 10$. Again, η has no effect on this particular scenario, because the cardinality of the basis is

864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917

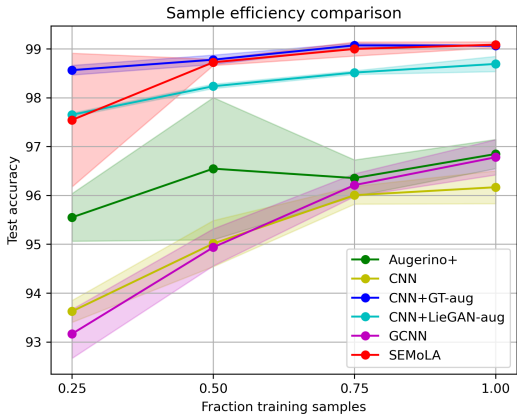


Figure 5: Comparison of the sample efficiency of SEMoLA, the base CNN model, Augerino+, GCNN, and the base CNN model trained with ground truth and LieGAN augmentations, which are trained with different fractions of the total training set samples of the RotatedMNIST dataset.

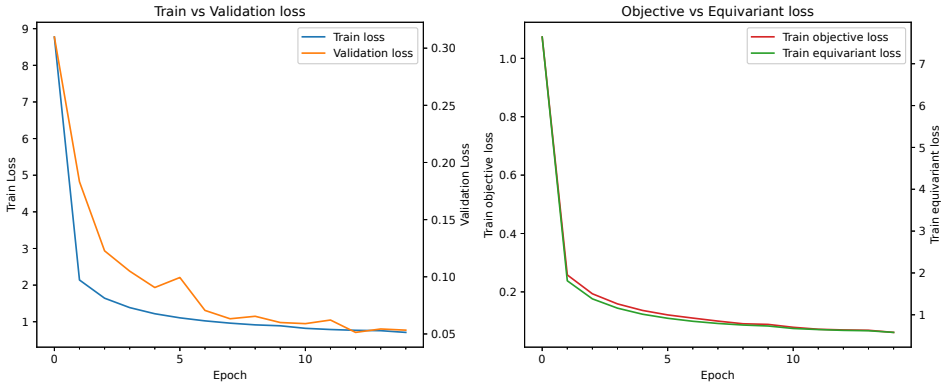


Figure 6: Loss curves of SEMoLA for RotatedMNIST dataset. (Left) Train and validation losses. (Right) Objective and equivariance losses.

$C = 1$. In terms of the dataset, we maintain the default train-test splits defined in Yang et al. (2023). We train all models for 100 epochs with a batch size of 64 for three runs with different random seeds and report the mean and standard deviation of the performances.

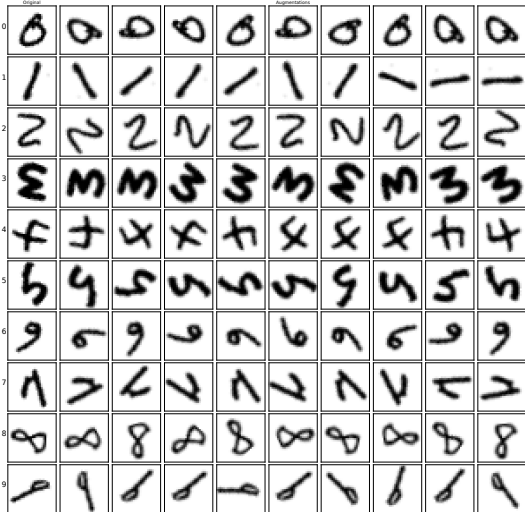
As mentioned in the paper, we test the impact of introducing different restrictions into the structure of the learned Lie algebra basis to explore the capabilities of the symmetry discovery approaches under varying degrees of auxiliary knowledge about the task. In that sense, we define different ‘masks’ that are multiplied by the learned Lie algebra before the default steps of the LieAugmenter in order to restrict the learnable entries being considered in the augmentations. As done in Yang et al. (2023), we consider a mask that implies searching for symmetries acting on the position and momentum of each mass separately (Mask_2), and a mask that allows interactions between only the positions and only the momenta across the different bodies (Mask_4). In addition, we again report the results from the paper in which we do not impose any mask (Mask_0). We add the same numerical suffixes to the considered models to indicate which mask each uses.

A.5.2 1-TIMESTEP PREDICTION

Firstly, we complement the experimental results presented in the paper for the case of predicting one output timestep from one input timestep.

In that sense, the symmetry discovery results presented in Figure 8 and 8 show that LieGAN (Yang et al., 2023) is able to identify the correct symmetry in the Mask_2 setting of the in-distribution

918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934



935 Figure 7: Example of augmentations generated by SEMoLA for the RotatedMNIST dataset. The
936 first column shows the original samples, and the subsequent ones correspond to the generated
937 augmentations.
938

939
940
941
942
943
944
945
946
947
948
949

scenario, as well as a noisy version of the alternative representation of the symmetry in the Mask_4 setting (see Appendix A.5.5), while Augerino+ (Yang et al., 2023) fails in both cases while producing much noisier results. In the out-of-distribution scenario, Augerino+ (Yang et al., 2023) still fails to learn the correct symmetry, while now LieGAN (Yang et al., 2023) fails to learn the symmetry in the Mask_2 setting. This result exemplifies the sensitivity of the GAN-based symmetry discovery methods to the uniformity of the considered dataset, since they rely on discovering distribution-preserving transformations. Meanwhile, we can see that SEMoLA successfully identifies the correct symmetry (or its alternative representation) under all masks while producing less noisy and sparser representations, which are generally preferable.

950
951
952
953

Table 8: Cosine and MAE similarities of the ground truth Lie algebra basis and the learned ones by each of the considered symmetry learning models in the in-distribution and out-of-distribution versions of the 2-body dataset with one input and output timestep. For each metric, we report the mean and standard deviation across three runs using different random seeds.

954
955
956
957
958
959
960
961
962
963
964
965
966
967
968

Model	In distribution		Out of distribution	
	Cosine	MAE	Cosine	MAE
Augerino+_0	0.2302±0.0122	1.2645±0.0900	0.1974±0.1417	1.0338±0.0353
Augerino+_2	0.3049±0.1762	0.3660±0.0790	0.2927±0.1684	0.3660±0.0782
Augerino+_4	0.6295±0.0071	0.4130±0.0914	0.6421±0.0043	0.7650±0.0114
LieGAN_0	0.2089±0.0289	0.3013±0.0185	0.0717±0.0280	0.2208±0.0223
LieGAN_2	0.9993±0.0002	0.1339±0.0222	0.6154±0.1573	0.1250±0.0003
LieGAN_4	0.6859±0.0259	0.2067±0.0517	0.7010±0.0013	0.2753±0.1045
SEMoLA_0	0.7096±0.0097	0.1873±0.0825	0.7046±0.0024	0.1293±0.0008
SEMoLA_2	1.0000±0.0000	0.0058±0.0008	0.6672±0.4706	0.0506±0.0619
SEMoLA_4	0.7075±0.0001	0.1836±0.0827	0.7071±0.0011	0.1263±0.0005

969
970
971

Moreover, we extend the results from Table 3 with the performance results for the different masking scenarios and their corresponding total train and test times together with the second equivariant error (Equation 4) in Tables 9 and 10, respectively. From these results, we can see that Augerino+ (Yang et al., 2023) displays the worst performance, due to it not being able to learn the correct symmetry,

972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

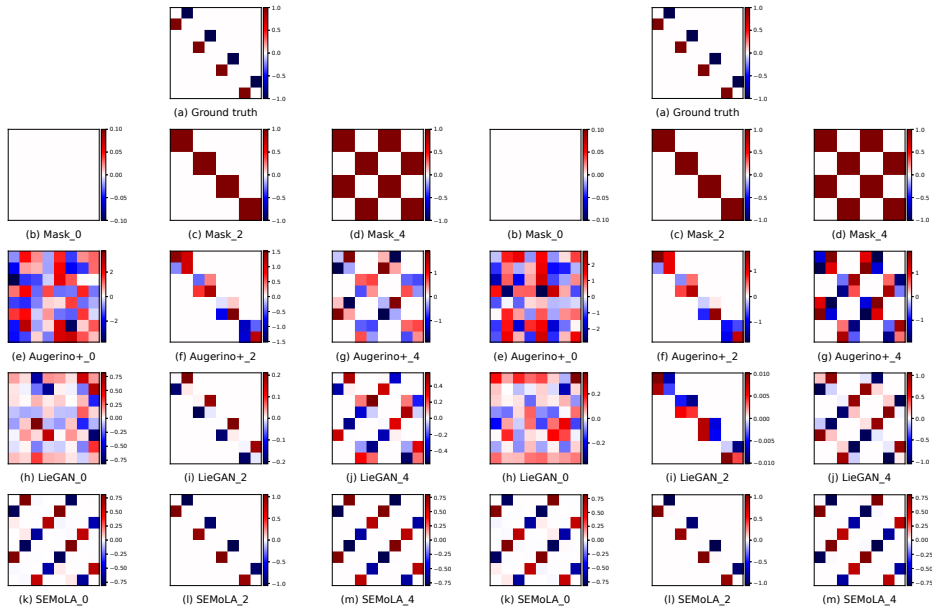


Figure 8: Comparison of the ground truth Lie algebra basis with those learned by the baselines LieGAN and Augerino+, and by our proposed SEMoLA approach for the in-distribution versions of the 2-body dataset (left) and the restricted out-of-distribution version (right) with one input and output timestep.

as previously discussed. The base MLP model has the second-worst performance, since it lacks any knowledge about the underlying equivariance of the data. Next, in terms of the MLP with the different fixed augmentations, we can see that using the ground truth symmetry provides the best performance among them, while using the symmetry discovered by LieGAN only performs well when it uses the Mask_4 setting. The EMLP (Finzi et al., 2021) model provides the best overall performance regardless of the Lie algebra basis considered for its definition, thanks to its ‘hard’ equivariance design, but notably it requires by far the longest train and test times. Lastly, SEMoLA provides the overall second-best performance regardless of the ‘mask’, being only worse than EMLP (Finzi et al., 2021) and performing very similarly to the MLP with ground truth augmentations. As a result, it provides the best performance among all prediction models that do not consider any a priori knowledge about the symmetry.

A.5.3 3-TIMESTEP PREDICTION

Next, we further extend the experimental results presented in the paper with the more complex scenario of predicting three output timesteps from three input timesteps.

We can find the symmetry discovery results in Figure 9 and 11, which are, for the most part, consistent with those previously shown in the 1-timestep case, but slightly noisier. Overall, while Augerino+ (Yang et al., 2023) and LieGAN (Yang et al., 2023) still struggle to learn the correct symmetry in most settings, SEMoLA is able to successfully learn it under all masks.

In terms of the performance results for this 3-timestep prediction setting, we can find them in Tables 12 and 13, where we can again observe very similar trends as in the 1-timestep prediction scenario. In particular, Augerino+ (Yang et al., 2023) and the MLP with LieGAN augmentations perform the worst, while the EMLP and the MLP with ground truth augmentations perform best, closely followed by SEMoLA, which provides the best performance among prediction models that do not consider any a priori knowledge about the symmetry.

Table 9: Results for the performance and equivariant error of the different considered models for in-distribution and out-of-distribution versions of the 2-body dataset with one input and output timestep. For each metric, we report the mean and standard deviation across three runs using different random seeds. The accuracies are expressed as percentages.

Model	In distribution		Out of distribution	
	MSE	Equiv. Error	MSE	Equiv. Error
Augerino+_0	5.28e+00±4.23e-01	1.90e-01±1.92e-03	3.26e+00±7.30e-01	1.94e-01±4.15e-03
Augerino+_2	9.67e+03±1.37e+04	2.37e-01±7.53e-02	1.48e+05±2.09e+05	1.95e-01±9.23e-03
Augerino+_4	1.79e+00±2.77e-02	1.87e-01±2.92e-04	1.55e+00±8.59e-03	1.86e-01±5.90e-04
MLP	5.16e-03±2.79e-03	2.48e-02±7.37e-03	1.86e-02±3.74e-03	3.33e-02±2.56e-03
MLP + GT aug	2.13e-05±1.56e-05	3.03e-03±1.03e-03	3.03e-05±8.81e-06	3.70e-03±1.14e-03
MLP + LieGAN_0 aug	3.23e-03±6.25e-04	1.91e-02±1.47e-03	1.15e-02±3.00e-03	2.84e-02±3.04e-03
MLP + LieGAN_2 aug	1.10e-03±6.98e-04	1.48e-02±4.80e-03	8.94e-03±9.31e-04	2.73e-02±1.10e-03
MLP + LieGAN_4 aug	2.85e-05±1.19e-05	3.02e-03±5.08e-04	5.12e-05±8.47e-06	2.78e-03±3.62e-04
EMLP + GT	1.74e-06±1.50e-06	5.32e-05±1.16e-05	4.64e-06±4.86e-06	5.27e-05±1.10e-05
EMLP + LieGAN_0	6.36e-06±4.56e-06	5.31e-05±1.13e-05	3.81e-06±2.62e-06	5.31e-05±1.09e-05
EMLP + LieGAN_2	4.82e-06±3.19e-06	5.33e-05±1.11e-05	1.93e-06±1.84e-06	5.30e-05±1.14e-05
EMLP + LieGAN_4	2.15e-06±1.98e-06	5.30e-05±1.16e-05	3.29e-06±2.33e-06	5.28e-05±1.14e-05
SEMoLA_0	1.39e-05±6.08e-07	2.78e-03±4.88e-04	2.38e-05±9.04e-06	2.59e-03±9.89e-05
SEMoLA_2	1.30e-05±5.58e-06	3.28e-03±9.36e-04	3.73e-05±2.97e-05	5.96e-03±2.82e-03
SEMoLA_4	1.45e-05±6.15e-06	3.15e-03±9.78e-04	1.61e-05±1.22e-06	3.13e-03±2.79e-04

Table 10: Results for the train and test times (MM:SS) and equivariant error from Equation 4 of the different considered models for in-distribution and out-of-distribution versions of the 2-body dataset with one input and output timestep. We report the mean of all the metrics and the standard deviation of the equivariant error across three runs using different random seeds.

Model	In distribution			Out of distribution		
	Train time	Test time	Equiv. Error'	Train time	Test time	Equiv. Error'
Augerino+_0	05 : 25	00 : 02	4.20e-01±1.85e-03	05 : 25	00 : 02	4.21e-01±6.31e-03
Augerino+_2	05 : 40	00 : 02	4.30e-01±5.04e-02	05 : 32	00 : 02	4.00e-01±6.40e-02
Augerino+_4	05 : 33	00 : 02	4.16e-01±1.64e-03	05 : 30	00 : 02	4.16e-01±3.58e-05
MLP	02 : 00	00 : 00	4.84e-02±1.58e-02	02 : 01	00 : 00	8.91e-02±6.24e-03
MLP + GT aug	02 : 48	00 : 00	6.35e-03±1.44e-03	03 : 10	00 : 00	6.53e-03±1.25e-03
MLP + LieGAN_0 aug	05 : 29	00 : 00	3.93e-02±1.73e-03	05 : 25	00 : 00	7.25e-02±1.01e-02
MLP + LieGAN_2 aug	05 : 19	00 : 00	2.83e-02±8.22e-03	05 : 15	00 : 00	6.68e-02±2.19e-03
MLP + LieGAN_4 aug	05 : 22	00 : 00	6.03e-03±8.00e-04	05 : 26	00 : 00	7.08e-03±1.67e-03
EMLP + GT	21 : 34	00 : 02	8.08e-05±1.76e-05	21 : 13	00 : 02	8.02e-05±1.72e-05
EMLP + LieGAN_0	19 : 22	00 : 02	8.06e-05±1.73e-05	15 : 39	00 : 02	8.08e-05±1.71e-05
EMLP + LieGAN_2	21 : 56	00 : 02	8.07e-05±1.72e-05	21 : 18	00 : 02	8.05e-05±1.75e-05
EMLP + LieGAN_4	21 : 36	00 : 02	8.04e-05±1.78e-05	16 : 09	00 : 02	8.03e-05±1.76e-05
SEMoLA_0	08 : 18	00 : 00	5.75e-03±2.26e-04	08 : 20	00 : 00	4.53e-03±2.52e-04
SEMoLA_2	08 : 28	00 : 00	5.34e-03±1.35e-03	08 : 26	00 : 00	9.61e-03±4.93e-03
SEMoLA_4	08 : 27	00 : 00	6.17e-03±1.61e-03	08 : 26	00 : 00	5.79e-03±7.28e-04

1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133

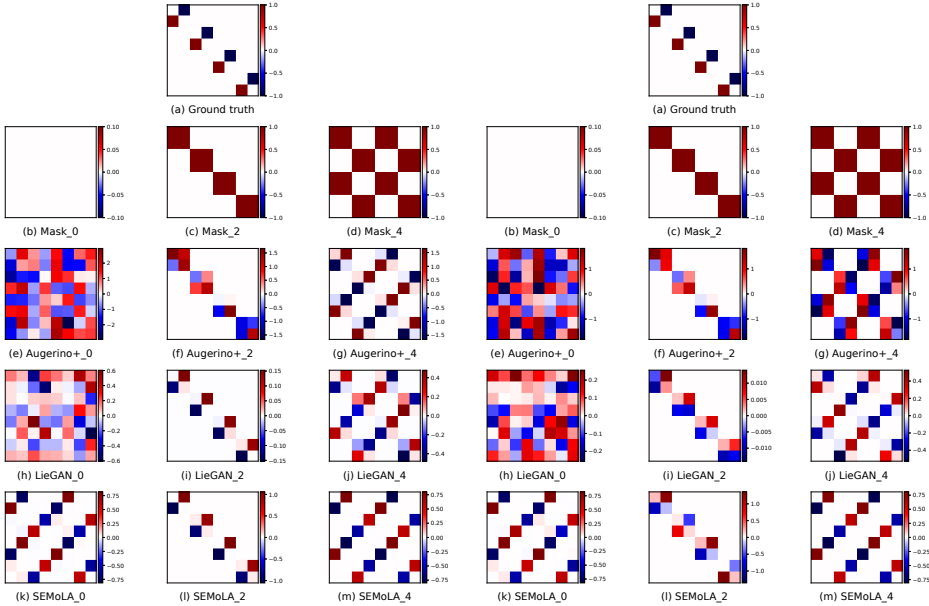


Figure 9: Comparison of the ground truth Lie algebra basis with those learned by the baselines LieGAN and Augerino+, and by our proposed SEMoLA approach for the in-distribution version of the 2-body dataset (left) and the restricted out-of-distribution version (right) with three input and output timesteps.

Table 11: Cosine and MAE similarities of the ground truth Lie algebra basis and the learned ones by each of the considered symmetry learning models in the in-distribution and out-of-distribution versions of the 2-body dataset with three input and output timesteps. For each metric, we report the mean and standard deviation across three runs using different random seeds.

Model	In distribution		Out of distribution	
	Cosine	MAE	Cosine	MAE
Augerino+_0	0.1927±0.0958	1.0683±0.1016	0.1127±0.1189	0.9469±0.0487
Augerino+_2	0.4402±0.2548	0.2905±0.0268	0.4001±0.3749	0.3239±0.0428
Augerino+_4	0.6679±0.0291	0.4709±0.1025	0.6131±0.0086	0.5870±0.0546
LieGAN_0	0.1943±0.0241	0.2734±0.0030	0.1317±0.0174	0.2149±0.0041
LieGAN_2	0.9991±0.0001	0.1189±0.0176	0.8732±0.0748	0.1251±0.0011
LieGAN_4	0.6841±0.0206	0.1729±0.0431	0.6857±0.0245	0.1360±0.0058
SEMoLA_0	0.7104±0.0008	0.1860±0.0829	0.5797±0.2021	0.1456±0.0302
SEMoLA_2	0.6665±0.4713	0.2114±0.0571	0.0017±0.0015	0.1482±0.0180
SEMoLA_4	0.7071±0.0006	0.3007±0.0003	0.7089±0.0001	0.1836±0.0828

Table 12: Results for the performance and equivariant error of the different considered models for in-distribution and out-of-distribution versions of the 2-body dataset with three input and output timesteps. For each metric, we report the mean and standard deviation across three runs using different random seeds. The accuracies are expressed as percentages.

Model	In distribution		Out of distribution	
	MSE	Equiv. Error	MSE	Equiv. Error
Augerino+_0	4.03e+00±3.49e-01	1.91e-01±2.17e-03	2.68e+00±4.94e-01	1.98e-01±9.13e-03
Augerino+_2	8.70e+00±1.00e+01	1.96e-01±8.81e-03	1.09e+01±1.33e+01	1.87e-01±8.77e-03
Augerino+_4	1.72e+00±1.03e-01	1.88e-01±1.83e-03	1.64e+00±7.75e-02	1.88e-01±2.73e-03
MLP	9.62e-03±4.32e-03	3.58e-02±7.57e-03	1.45e-02±1.30e-03	3.51e-02±1.39e-03
MLP + GT aug	3.96e-05±1.75e-05	5.08e-03±1.19e-03	4.92e-05±2.25e-05	3.88e-03±6.66e-04
MLP + LieGAN_0 aug	8.83e-03±2.72e-03	3.59e-02±5.83e-03	1.88e-02±4.80e-03	4.15e-02±5.98e-03
MLP + LieGAN_2 aug	2.71e-03±2.39e-04	2.26e-02±1.16e-03	1.83e-02±4.37e-03	4.05e-02±3.98e-03
MLP + LieGAN_4 aug	8.86e-05±7.58e-05	4.24e-03±1.10e-03	7.94e-05±2.55e-05	4.78e-03±7.64e-04
EMLP + GT	4.54e-06±2.20e-06	2.91e-05±4.71e-06	6.30e-06±3.68e-06	2.90e-05±4.79e-06
EMLP + LieGAN_0	2.99e-06±1.83e-06	2.91e-05±4.85e-06	5.63e-06±2.62e-06	2.91e-05±4.62e-06
EMLP + LieGAN_2	2.94e-06±1.21e-06	2.91e-05±4.75e-06	8.04e-06±5.55e-06	2.92e-05±4.66e-06
EMLP + LieGAN_4	3.26e-06±1.72e-06	2.92e-05±4.77e-06	6.40e-06±6.40e-06	2.92e-05±4.68e-06
SEMoLA_0	5.98e-05±2.17e-05	4.07e-03±1.50e-03	1.59e-04±1.93e-04	4.90e-03±1.88e-03
SEMoLA_2	1.75e-05±6.22e-06	4.58e-03±4.07e-04	3.67e-04±1.10e-04	7.97e-03±1.45e-03
SEMoLA_4	2.85e-05±3.92e-06	3.97e-03±6.82e-04	5.32e-05±1.28e-05	4.22e-03±1.75e-04

Table 13: Results for the train and test times (MM:SS) and equivariant error from Equation 4 of the different considered models for in-distribution and out-of-distribution versions of the 2-body dataset with three input and output timesteps. We report the mean of all the metrics and the standard deviation of the equivariant error across three runs using different random seeds.

Model	In distribution			Out of distribution		
	Train time	Test time	Equiv. Error'	Train time	Test time	Equiv. Error'
Augerino+_0	04 : 20	00 : 01	4.22e-01±3.07e-03	04 : 20	00 : 01	4.22e-01±8.34e-03
Augerino+_2	04 : 20	00 : 01	4.13e-01±1.26e-02	04 : 23	00 : 01	4.13e-01±1.53e-03
Augerino+_4	04 : 21	00 : 01	4.19e-01±4.20e-03	04 : 25	00 : 01	4.20e-01±3.19e-03
MLP	01 : 36	00 : 00	7.17e-02±1.53e-02	01 : 32	00 : 00	8.36e-02±3.20e-03
MLP + GT aug	02 : 34	00 : 00	1.10e-02±3.29e-03	02 : 33	00 : 00	8.62e-03±8.36e-04
MLP + LieGAN_0 aug	04 : 22	00 : 00	7.20e-02±1.13e-02	04 : 21	00 : 00	9.55e-02±1.45e-02
MLP + LieGAN_2 aug	04 : 16	00 : 00	4.36e-02±2.40e-03	04 : 16	00 : 00	9.53e-02±1.06e-02
MLP + LieGAN_4 aug	04 : 17	00 : 00	8.10e-03±1.21e-03	04 : 13	00 : 00	9.62e-03±3.32e-03
EMLP + GT	17 : 32	00 : 02	4.47e-05±7.24e-06	16 : 55	00 : 02	4.46e-05±7.31e-06
EMLP + LieGAN_0	16 : 57	00 : 02	4.48e-05±7.46e-06	17 : 10	00 : 02	4.47e-05±7.16e-06
EMLP + LieGAN_2	17 : 05	00 : 02	4.47e-05±7.33e-06	17 : 26	00 : 02	4.47e-05±7.19e-06
EMLP + LieGAN_4	16 : 50	00 : 02	4.49e-05±7.39e-06	15 : 15	00 : 02	4.47e-05±7.20e-06
SEMoLA_0	06 : 37	00 : 00	7.62e-03±1.95e-03	06 : 35	00 : 00	1.04e-02±4.38e-03
SEMoLA_2	06 : 43	00 : 00	8.50e-03±8.16e-04	06 : 40	00 : 00	2.13e-02±2.66e-03
SEMoLA_4	06 : 46	00 : 00	8.37e-03±9.29e-04	06 : 48	00 : 00	8.11e-03±3.96e-04

A.5.4 EXAMPLE SAMPLED AUGMENTATIONS

We again illustrate some examples of the augmentations generated by SEMoLA for a particular input sample in Figure 10, where we can see that both the input and its corresponding label are properly rotated as corresponds to the learned rotational equivariance.

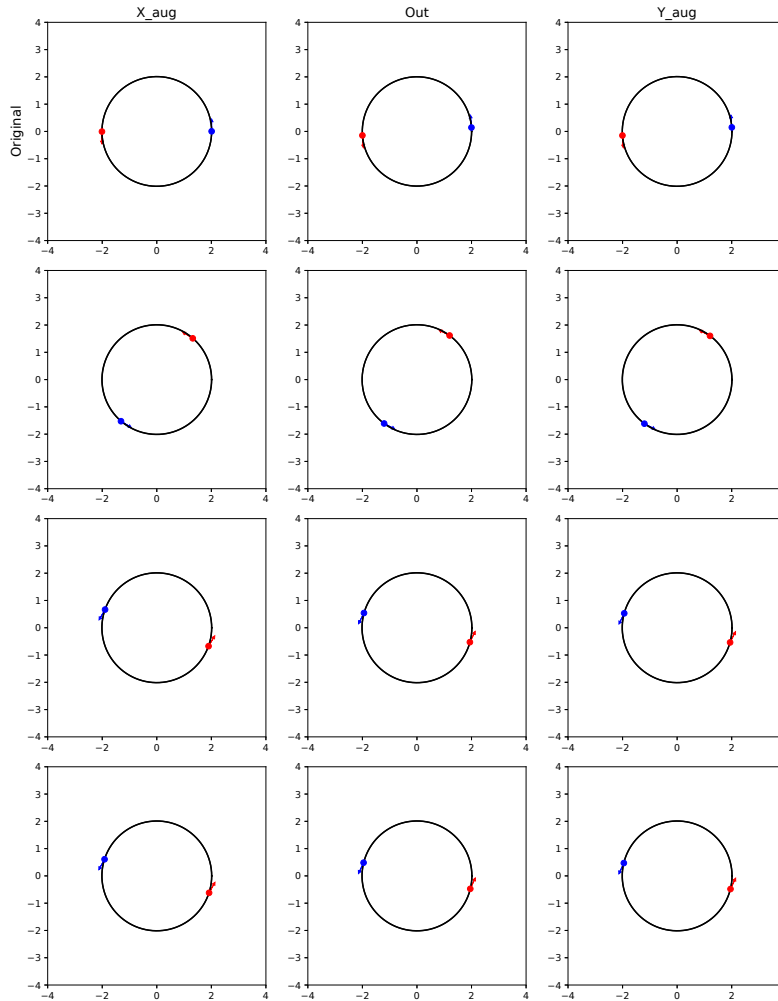


Figure 10: Examples of augmentations generated by SEMoLA for the 2-body dataset. The top row shows (from left to right) the original input sample, the output of the model for that sample, and the target prediction. The rows below show different augmented versions.

A.5.5 ALTERNATIVE REPRESENTATION OF THE ROTATION SYMMETRY

Throughout our experiments in this dataset, we have observed what we claim to be an alternative representation of the correct symmetry in the context of this 2-body dataset. In that sense, following Yang et al. (2023), we present the derivation that confirms that claim.

1242 Firstly, we can write this alternative representation of the Lie algebra basis in the following way to
 1243 highlight its structure after removing the slightly noisy entries:
 1244

$$1245 \quad R = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

$$1246 \quad L = \begin{bmatrix} R & & -R & \\ & R & & -R \\ -R & & R & \\ & -R & & R \end{bmatrix}$$

1251 Then, we can express the result of applying the matrix exponential to a weighted version of that basis
 1252 as follows:
 1253

$$1254 \quad \exp(wL) = I + \begin{bmatrix} G(w) & & -G(w) & \\ & G(w) & & -G(w) \\ -G(w) & & G(w) & \\ & -G(w) & & G(w) \end{bmatrix}$$

$$1255 \quad G(w) = \sum_{n=0}^{\infty} \frac{(-1)^n 2^{2n} w^{2n+1}}{(2n+1)!} R + \sum_{n=1}^{\infty} \frac{(-1)^n 2^{2n-1} w^{2n}}{(2n)!} I$$

1263 As a result, we can define its action on the input data as follows:
 1264

$$1265 \quad \exp(wL) \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{p}_1 \\ \mathbf{q}_2 \\ \mathbf{p}_2 \end{bmatrix} = \text{diag}(I + 2G(w)) \begin{bmatrix} \mathbf{q}_1 \\ \mathbf{p}_1 \\ \mathbf{q}_2 \\ \mathbf{p}_2 \end{bmatrix},$$

1269 where we have used our knowledge about the dataset, i.e., the origin is at the center of mass, and we
 1270 have $m_1 = m_2$, $\mathbf{q}_1 = -\mathbf{q}_2$, and $\mathbf{p}_1 = -\mathbf{p}_2$.

1271 Lastly, we can notice the following relation between the resulting expression and the ground truth
 1272 rotation symmetry:
 1273

$$1274 \quad I + 2G(w) = \sum_{n=0}^{\infty} \frac{(-1)^n 2^{2n+1} w^{2n+1}}{(2n+1)!} R + \sum_{n=0}^{\infty} \frac{(-1)^n 2^{2n} w^{2n}}{(2n)!} I$$

$$1275 \quad = \begin{bmatrix} \cos 2w & -\sin 2w \\ \sin 2w & \cos 2w \end{bmatrix}$$

1281 A.6 QM9

1282 For our experiments using the QM9 dataset, we set the following hyperparameter values: $\alpha = 1.0$,
 1283 $\beta = 3.0$, $\lambda = 0.1$, $\eta = 1.0$, $\nu = 0.001$, $\gamma = 3.0$ for the uniform distribution in the LieAugmenter,
 1284 and number of augmentations $K = 3$. In addition, we maintain the default train-test splits of the
 1285 dataset (Blum and Reymond, 2009; Rupp et al., 2012). We train SEMoLA for 500 epochs with a
 1286 batch size of 125. We should note that the reported results from Benton et al. (2020) and Yang et al.
 1287 (2023) used a slightly different setup, e.g., with the batch size being 75 and the inclusion of a learning
 1288 rate scheduling strategy. Nevertheless, we believe that, despite those differences, the obtained results
 1289 are comparable.

1290 In order to complement the results presented in the paper (Table 4), we also report the train and test
 1291 times and the second equivariant error in Table 14.
 1292

1293 In addition, we build on our previous qualitative analysis of the correctness of the learned Lie algebra
 1294 bases by also computing a quantitative metric that takes into account the mismatched position of
 1295 the elements of the ground truth and learned bases. In that sense, we compute the average cosine
 similarity across basis elements between the ground truth Lie algebra basis and different permutations

Table 14: Results for the time (HH:MM:SS) and equivariant error from Equation 4 of SEMoLA for the different considered prediction targets of the QM9 dataset.

Model	HOMO			LUMO		
	Train time	Test time	Equiv. Error'	Train time	Test time	Equiv. Error'
SEMoLA	15 : 17 : 12	00 : 00 : 04	0.0045	15 : 19 : 15	00 : 00 : 04	0.0019

of the elements of the learned basis, and report the permutation that matches the elements in the ground truth basis in a way that maximizes the similarity for each of the prediction tasks in Table 15. As we can see, the learned Lie algebra bases exactly match the ground truth for this metric after being permuted for both prediction tasks.

Table 15: Maximum average cosine similarity and its corresponding permutation of the elements of the learned Lie algebra basis for each of the considered prediction tasks in the QM9 dataset.

Maximum average cosine similarity (Permutation)	
HOMO	1.0000 (3, 1, 0, 2, 5, 4)
LUMO	1.0000 (4, 0, 1, 5, 2, 3)

Lastly, we also carry out an empirical analysis of the convergence of the model for this more complex task through the loss curves shown in Figure 11, where we can again observe the quick convergence of both the equivariance and objective losses despite the complexity of the associated optimization problem.

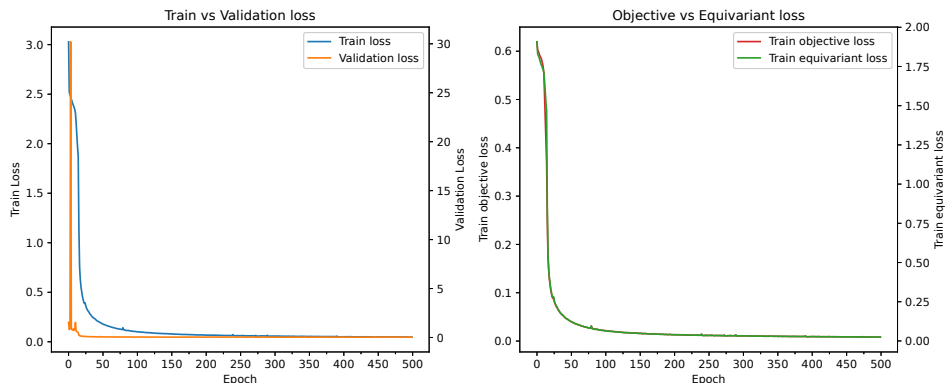


Figure 11: Loss curves for LUMO task of the QM9 dataset. (Left) Train and validation losses. (Right) Objective and equivariance losses.

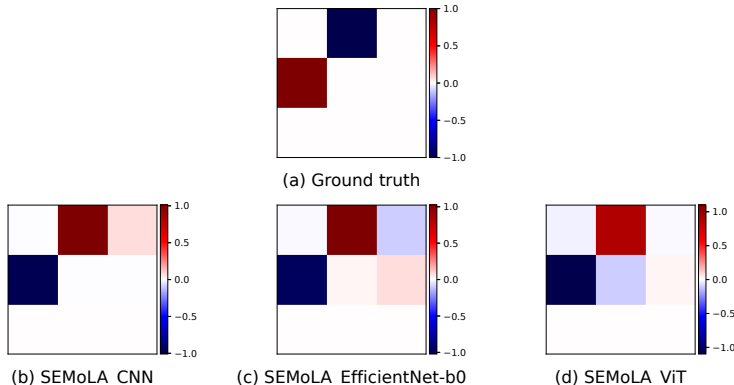
A.7 COLORECTAL CANCER DETECTION

As another more complex scenario for analyzing the performance of our proposal, we consider the task of human colorectal cancer (CRC) detection from stained histological images of human tissues. In particular, we consider the datasets NCT-CRC-HE-100K for training and CRC-VAL-HE-7K for testing (Kather et al., 2018), each containing 100,000 and 7,000 samples, respectively, of histological slices of human colorectal cancer and normal tissue, represented as 224x224 RGB images. Rotational invariance has been shown to be particularly important for this classification task (Gerken and Kessel, 2024).

The main goal of this experiment is to explore the applicability of our approach to a more complex and relevant computer vision scenario, where the images are larger and in color. In addition, we also seek to study the effect of applying SEMoLA to more complex architectures besides a simple CNN, e.g., Vision Transformer (ViT) (Dosovitskiy et al., 2020) and EfficientNet (Tan and Le, 2019).

1350 We should note that in this experiment, we do not seek to attain the highest possible performance for
 1351 the considered models in this dataset, but rather observe the effect of applying SEMoLA in contrast
 1352 to using ground truth augmentations on each of the base models. As a result, we do not tune the
 1353 hyperparameters, which we set similarly to previous experiments: $\alpha = 1.0$, $\beta = 7.0$, $\lambda = 0.1$,
 1354 $\eta = 0.0$, $\nu = 0.01$, $\gamma = 3.0$ for the uniform distribution in the LieAugmenter, and number of
 1355 augmentations $K = 10$. We train all models for 25 epochs with a batch size of 64.

1356 In terms of symmetry discovery, we follow the same procedure as for the RotatedMNIST dataset and
 1357 set the cardinality of the Lie algebra basis to $C = 1$ while searching over the six-dimensional space
 1358 of 2D affine transformations, which are defined to act on the pixel coordinates of the images. With
 1359 that definition, SEMoLA is able to learn the correct symmetry for all of the considered architectures,
 1360 as shown in Figure 12.



1376 Figure 12: Comparison of the ground truth Lie algebra basis with those learned by the SEMoLA for
 1377 each of the considered base models in the CRC dataset.

1379 As we can see in the results in Table 16, SEMoLA provides an equal or better performance than using
 1380 ground truth augmentations for all three models. This observation coincides with and reinforces our
 1381 previous results, demonstrating once more the benefits of our approach.

1384 Table 16: Results for the performance and equivariant error of the different models considered in the
 1385 CRC dataset. For each metric, we report the mean and standard deviation across three runs using
 1386 different random seeds. The accuracies are expressed as percentages.

Model	Accuracy	Equiv. Error
CNN	84.87±4.44	2.95±0.15
CNN + GT aug	85.27±3.34	1.75±0.29
SEMoLA (CNN)	89.42±0.89	1.86±0.21
EfficientNet-b0	93.04±1.18	2.55±0.24
EfficientNet-b0 + GT aug	93.35±1.93	1.49±0.14
SEMoLA (EfficientNet-b0)	92.74±2.55	1.21±0.10
ViT	75.69±3.90	2.02±0.42
ViT + GT aug	81.16±2.67	1.09±0.07
SEMoLA (ViT)	85.63±2.93	0.82±0.15

1402 Finally, we also show some examples of the augmentations generated by SEMoLA for each label of
 1403 the CRC dataset in Figure 13.

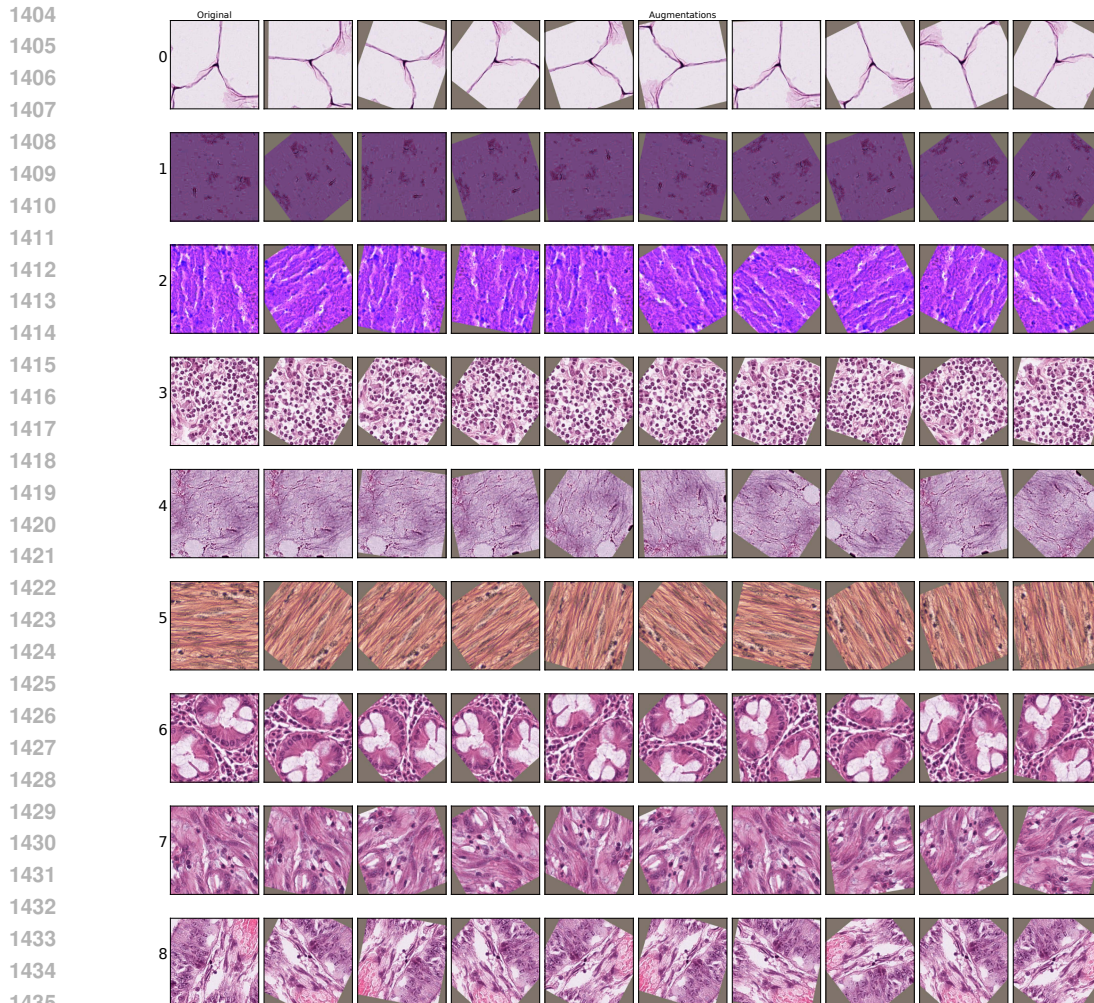


Figure 13: Example of augmentations generated by SEMoLA for the CRC dataset. The first column shows the original samples, and the subsequent ones correspond to the generated augmentations.

A.8 SYNTHETIC DATASETS

Next, we also explored the ability of SEMoLA to identify the correct Lie group when the symmetry of the data corresponds to discrete subgroups of a Lie group instead of continuous groups, as we explored in the main experiments. To that effect, we construct two synthetic regression datasets based on some scenarios explored in Yang et al. (2023).

For both of these datasets, we set the following hyperparameter values for our model: $\alpha = 1.0$, $\beta = 5.0$, $\lambda = 0.1$, $\eta = 0.0$, $\nu = 0.001$, $\gamma = 3.0$ for the uniform distribution in the LieAugmenter, and number of augmentations $K = 10$. Once more, η has no effect on this particular scenario, because the cardinality of the bases is $C = 1$. We generate 54,000 samples for the training set, 6,000 for the validation set, and 10,000 for the test set.

In order to attempt to model the discrete nature of the studied groups, we modify the sampling strategy of the coefficients of the linear combination of the elements of the Lie algebra basis by sampling from a uniform integer grid instead of a uniform distribution. Since prior knowledge about the discrete or continuous nature of the underlying group is not always readily available, we also explore the performance of SEMoLA with the previous uniform sampling strategy to show its robustness to such a possible misspecification. We distinguish the results by the nomenclature ‘SEMoLA (discrete)’ and ‘SEMoLA (continuous)’. Overall, we observe that both sampling strategies provide similarly strong

performance (especially for symmetry discovery), with the discrete formulation providing slightly higher task performance in our experiments.

A.8.1 DISCRETE ROTATION

As the first synthetic dataset, we consider the synthetic regression problem given by $f(x, y, z) = z / (1 + \arctan \frac{y}{x} \bmod \frac{2\pi}{k})$, which is invariant to rotations of multiples of $2\pi/k$ in the xy plane, thus corresponding to a discrete cyclic subgroup of $SO(2)$ with size k . Specifically, we consider the case where $k = 6$ and $x, y, z \sim \mathcal{N}(0, I)$.

From Figure 14 and Table 17, we can see that SEMoLA is capable of learning the correct Lie group for the underlying discrete symmetry, while Augerino+(Yang et al., 2023) fails and LieGAN (Yang et al., 2023) provides a much noisier representation.

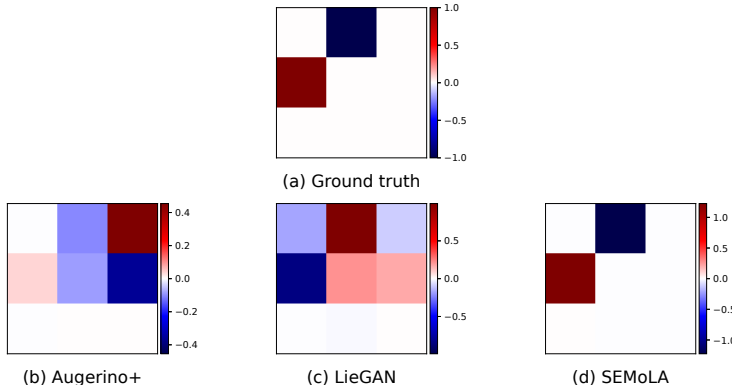


Figure 14: Comparison of the ground truth Lie algebra basis with those learned by the baselines Augerino+ and LieGAN and our proposed SEMoLA (continuous) in the DiscreteRotation6 dataset.

Table 17: Cosine and MAE similarities of the ground truth Lie algebra basis with those learned by each of the considered symmetry learning models in the DiscreteRotation6 dataset. For each metric, we report the mean and standard deviation across three runs using different random seeds.

In distribution		
Model	Cosine	MAE
Augerino+	0.4232±0.4080	0.3863±0.0581
LieGAN	0.6577±0.0095	0.4419±0.0432
SEMoLA (continuous)	1.0000±0.0000	0.1989±0.2092
SEMoLA (discrete)	0.8588±0.0017	0.2998±0.2098

In terms of the performance of the different models, we can see in Tables 18 and 19 that SEMoLA (continuous), the MLP with ground truth and LieGAN augmentations, and Augerino+(Yang et al., 2023) all provide extremely similar performance, while the base MLP performs best due to this being a very simple problem for which encoding invariance is not strictly necessary, and SEMoLA (discrete) provides the second best performance.

A.8.2 PARTIAL PERMUTATION

Next, we seek to show that SEMoLA is able to learn symmetry groups beyond those concerning rotations. In particular, we consider the function $f(x) = x_1 + x_2 + x_3 + x_4^2 - x_5^2$, $x \in \mathbb{R}^5$, which is invariant to partial permutations, with the output remaining the same if we permute the first three dimensions of the input, but changes if we permute the last two. Specifically, we sample $x \sim \mathcal{U}[-100, 100]$.

Table 18: Results for the performance and equivariant error of the different considered models in the DiscreteRotation6 dataset. For each metric, we report the mean and standard deviation across three runs using different random seeds.

In distribution		
Model	MSE	Equiv. Error
Augerino+	2.09e-02±5.56e-04	1.08e-02±4.22e-03
MLP	4.06e-03±3.02e-04	9.57e-02±1.54e-03
MLP + GT aug	2.10e-02±1.19e-03	2.31e-03±2.36e-04
MLP + LieGAN aug	2.10e-02±1.10e-03	3.28e-03±3.45e-04
SEMoLA (continuous)	2.12e-02±1.24e-03	2.46e-03±3.29e-04
SEMoLA (discrete)	3.36e-03±4.56e-04	3.09e-02±1.80e-03

Table 19: Results for the train and test times (MM:SS) and equivariant error from Equation 4 of the different considered models for the DiscreteRotation6 dataset. We report the mean of all the metrics and the standard deviation of the equivariant error across three runs using different random seeds.

In distribution			
Model	Train time	Test time	Equiv. Error'
Augerino+	00 : 09 : 29	00 : 00 : 01	1.36e-02±5.32e-03
MLP	00 : 03 : 31	00 : 00 : 00	1.24e-01±2.39e-03
MLP + GT aug	00 : 05 : 35	00 : 00 : 00	2.91e-03±2.35e-04
MLP + LieGAN aug	00 : 09 : 21	00 : 00 : 00	4.21e-03±4.20e-04
SEMoLA (continuous)	00 : 14 : 24	00 : 00 : 00	3.18e-03±4.07e-04
SEMoLA (discrete)	00 : 14 : 51	00 : 00 : 00	4.41e-02±2.06e-03

From Figure 15 and Table 20, we can see that SEMoLA is capable of learning the correct Lie group for the underlying discrete symmetry, while Augerino+(Yang et al., 2023) and LieGAN (Yang et al., 2023) both fail. We should note, however, that in Yang et al. (2023), the authors report that LieGAN (Yang et al., 2023) is able to learn this symmetry, but since the exact hyperparameter values they used are not explicitly stated, we could not reproduce those results exactly.

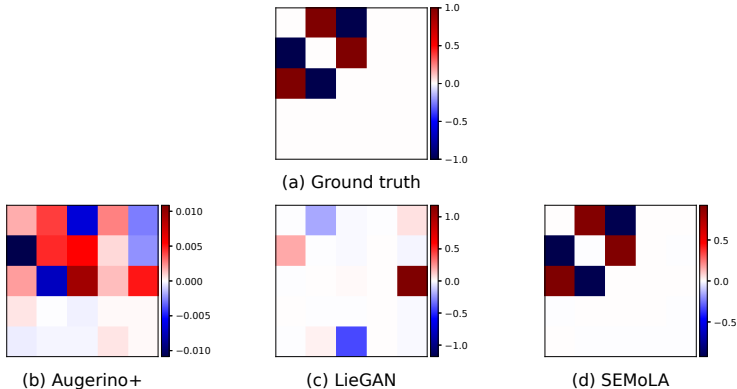


Figure 15: Comparison of the ground truth Lie algebra basis with those learned by the baselines Augerino+ and LieGAN and our proposed SEMoLA (continuous) in the PartialPermutation dataset.

Table 20: Cosine and MAE similarities of the ground truth Lie algebra basis with those learned by each of the considered symmetry learning models in the PartialPermutation dataset. For each metric, we report the mean and standard deviation across three runs using different random seeds.

In distribution		
Model	Cosine	MAE
Augerino+	0.2890±0.3236	0.2411±0.0009
LieGAN	0.1818±0.1255	0.3315±0.0024
SEMoLA (continuous)	0.9998±0.0002	0.3140±0.2069
SEMoLA (discrete)	0.9987±0.0006	0.3170±0.2045

In terms of the performance of the different models, we can see in Tables 21 and 22 that SEMoLA and the MLP with ground truth augmentations perform best, while the other approaches all produce considerably higher errors.

Table 21: Results for the performance and equivariant error of the different considered models for the PartialPermutation dataset. For each metric, we report the mean and standard deviation across three runs using different random seeds. The accuracies are expressed as percentages.

In distribution		
Model	MSE	Equiv. Error
Augerino+	3.38e+03±8.96e+02	2.52e+01±4.52e+00
MLP	1.97e+03±1.20e+03	1.23e+01±3.21e+00
MLP + GT aug	1.69e+03±8.44e+02	1.41e+01±2.06e+00
MLP + LieGAN aug	1.24e+04±2.90e+03	1.05e+01±2.12e+00
SEMoLA (continuous)	1.77e+03±3.57e+02	1.37e+01±4.55e+00
SEMoLA (discrete)	1.49e+03±6.55e+01	1.19e+01±2.76e+00

Table 22: Results for the train and test times (MM:SS) and equivariant error from Equation 4 of the different considered models for the PartialPermutation dataset. We report the mean of all the metrics and the standard deviation of the equivariant error across three runs using different random seeds.

In distribution			
Model	Train time	Test time	Equiv. Error'
Augerino+	00 : 09 : 55	00 : 00 : 01	3.20e+01±5.35e+00
MLP	00 : 03 : 40	00 : 00 : 00	1.58e+01±4.25e+00
MLP + GT aug	00 : 05 : 33	00 : 00 : 00	1.80e+01±2.68e+00
MLP + LieGAN aug	00 : 09 : 54	00 : 00 : 00	1.34e+01±2.59e+00
SEMoLA (continuous)	00 : 15 : 09	00 : 00 : 00	1.75e+01±5.80e+00
SEMoLA (discrete)	00 : 15 : 04	00 : 00 : 00	1.49e+01±3.45e+00

A.9 ABLATIONS

Lastly, we perform a series of ablations of the different loss weights and other hyperparameters of SEMoLA in order to show their influence on the performance. To that effect, we again consider the RotatedMNIST dataset, which has the fastest execution time among the considered datasets, thus allowing for a more extensive analysis.

From Figure 16 (left), we observe that the performance of our model is quite robust to changes in the values of the weights for the different terms of the loss function except for β , i.e., the weight for the \mathcal{L}_{equiv} term. In particular, decreasing the value of β leads to a degraded performance, as expected, since the LieAugmenter will not receive a strong enough learning signal to properly learn the correct symmetry. In addition, in Figure 16 (right), we can see that the number of sampled augmentations per input sample can also considerably affect the performance of the model, although not so drastically. In particular, increasing the number of augmentations generally leads to an improved performance, as we could expect, but after a certain point, the performance starts decreasing and becomes more variable across runs. One possible reason for that behavior is that at that point the number of augmented samples overcomes the importance of the original samples (due to the number of augmentations per sample being a fourth of the batch size), and the learning behavior becomes impaired due to the model focusing too much on the augmented samples over the original ones.

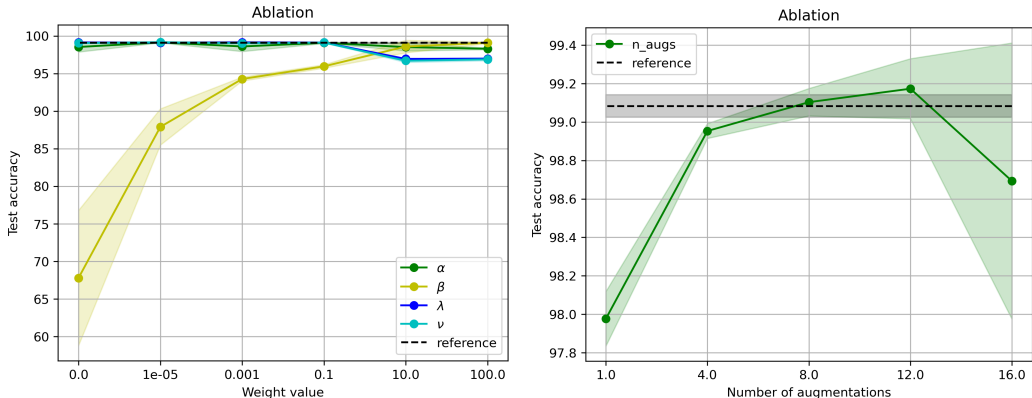


Figure 16: Ablation results for the accuracy metric in the RotatedMNIST dataset for varying values for the weight terms of the loss function (left) and the number of augmentations per sample (right). The reference results correspond to the hyperparameters considered in the main experiments. All results are reported as the mean and standard deviation over three runs with different random seeds.

In Figure 17, we can find the corresponding results for the ablation analysis focusing on the cosine similarity of the learned Lie algebra basis to the ground truth instead of the performance metric. For the weight terms of the loss function, we can see that modifying α (the weight of \mathcal{L}_{obj}) has little effect on the symmetry discovery process. Modifying λ and ν also has little impact below a threshold of 0.1, but when those terms are increased too much, they lead to a degraded symmetry discovery performance, as the difference of the augmentations from the original data (λ) or the sparsity of the learned basis (ν) become too prominent in the loss. Furthermore, β also has a relevant impact on the symmetry discovery process, where we observe that higher values of that weight help the performance of the model. Lastly, the number of augmentations per sample has little impact on the symmetry discovery process for values below 16, but at that point the performance degrades, possibly for the same reason that we hypothesized for the performance metric.

Next, we also explore the effect of the chosen range γ for the uniform sampling of coefficients for the linear combination of elements of the Lie algebra basis, which is then mapped to a group element. This is a hyperparameter that we did not extensively tune in our experiments, and its values were chosen mainly seeking to encourage the sampling of diverse elements of the group while not being excessively large to avoid distortions in the uniformity of the sampling (e.g., for rotation groups the sampled elements would “wrap around” for larger ranges, thus distorting the effective sampling density of the group elements).

In order to verify our intuition and solidify our explanation in that regard, we run an additional ablation experiment for the γ hyperparameter using the RotatedMNIST dataset, the results of which are shown in Table 23. In these results, we would start observing the “wrap around” effect starting at $\gamma = 5.0$ if the model were to learn the ground truth symmetry, which may explain the decrease in performance and cosine similarity for that and the larger values of this hyperparameter. In the case of $\gamma = 13.0$, we hypothesize that the slight increase in the performance and similarity of the basis is

1674
 1675
 1676
 1677
 1678
 1679
 1680
 1681
 1682
 1683
 1684
 1685
 1686
 1687
 1688
 1689
 1690
 1691
 1692
 1693
 1694
 1695
 1696
 1697
 1698
 1699
 1700
 1701
 1702
 1703
 1704
 1705
 1706
 1707
 1708
 1709
 1710
 1711
 1712
 1713
 1714
 1715
 1716
 1717
 1718
 1719
 1720
 1721
 1722
 1723
 1724
 1725
 1726
 1727

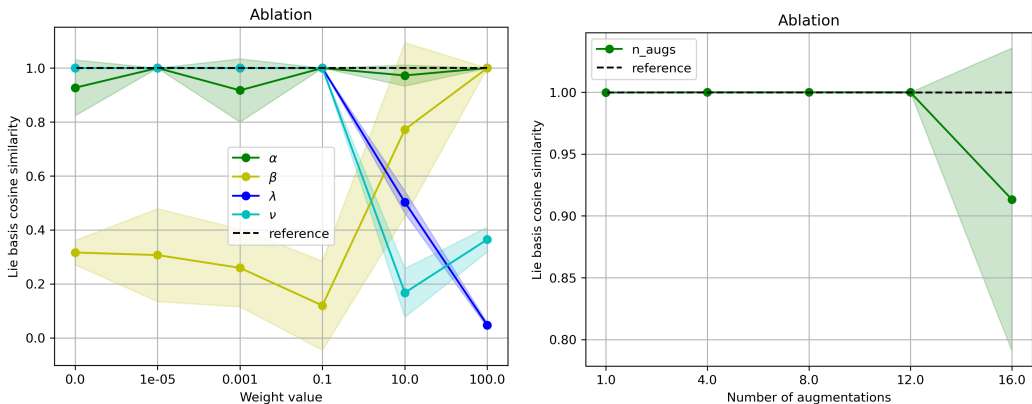


Figure 17: Ablation results for the cosine similarity of the learned Lie algebra basis to the ground truth in the RotatedMNIST dataset for varying values for the weight terms of the loss function (left) and the number of augmentations per sample (right). The reference results correspond to the hyperparameters considered in the main experiments. All results are reported as the mean and standard deviation over three runs with different random seeds.

due to the fact that learning the ground truth Lie algebra basis would lead to rotations approximately in the range $[-2\pi, 2\pi]$, thus circumventing the issue.

Table 23: Results for the ablation of γ . We report the mean and standard deviation for both metrics across three runs using different random seeds. The accuracies are expressed as percentages.

γ	Test accuracy	Cosine similarity Lie algebra basis
1.0	97.73±0.17	0.9999±0.0001
2.0	98.04±0.22	1.0000±0.0000
3.0	99.08±0.06	0.9997±0.0002
5.0	95.97±3.24	0.7860±0.3026
8.0	93.11±3.79	0.5706±0.3571
9.0	94.93±4.94	0.6783±0.4549
10.0	92.56±4.19	0.6273±0.2656
11.0	89.28±2.96	0.3709±0.2551
12.0	96.74±2.14	0.6881±0.4381
13.0	96.98±0.72	0.9812±0.0178
14.0	93.32±2.99	0.7236±0.2095
15.0	91.89±2.20	0.5346±0.0607
16.0	93.52±0.91	0.2104±0.1597
17.0	90.28±3.33	0.2779±0.1739
21.0	89.00±4.13	0.0006±0.0006

Finally, we also explore the symmetry discovery performance of SEMoLA under a misspecified Lie algebra basis cardinality C in Figure 18. As we can see, under the slight misspecification of setting $C = 2$, we observe redundancy across the different elements of the learned basis, which shows that our model is able to robustly discover the correct basis. In the more extreme cases of setting $C = 4$ or $C = 6$, we start observing a slight degradation in the discovered basis, which is somewhat noisier. In addition, while some elements in the basis are inverses of one another, their linear combination would still lead to a valid rotation augmentation. Therefore, while knowing the correct (or a close

approximation) of the basis cardinality a priori helps the symmetry discovery performance of the model, we have shown that it can be robust in scenarios where that knowledge may not be available.

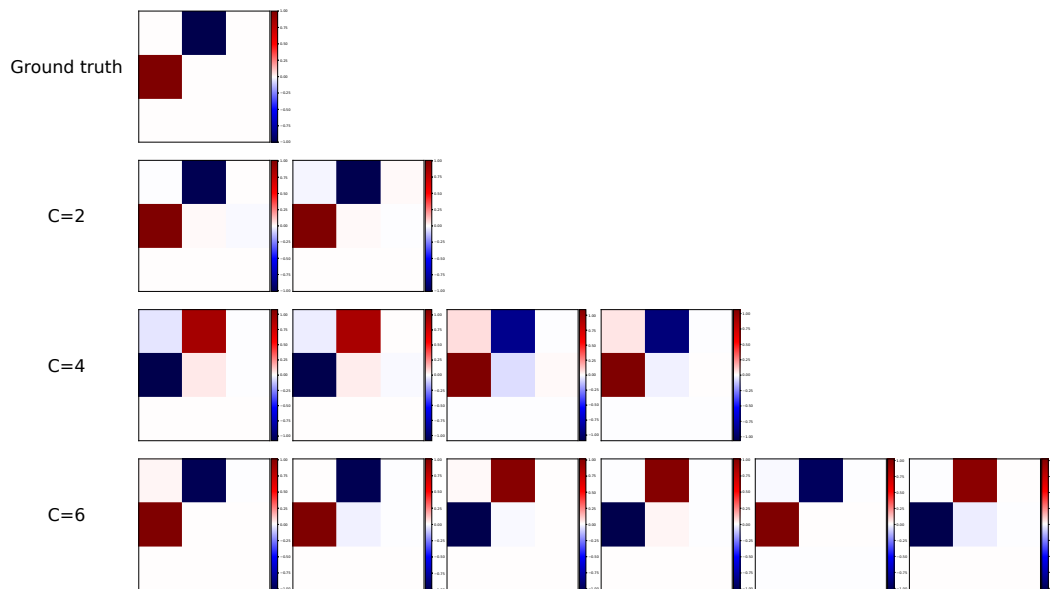


Figure 18: SEMoLA is able to correctly discover a representation of the underlying symmetry even when the cardinality of the Lie algebra basis is misspecified in the RotatedMNIST dataset.

A.10 DATASET WITHOUT SYMMETRIES

As an additional experiment, we investigate the behavior of SEMoLA when applied to a dataset that lacks any symmetry in order to determine whether it learns an arbitrary spurious symmetry or if, on the contrary, it discovers a representation that can be easily used to diagnose the lack of symmetry in the dataset and that does not negatively impact the performance of the method.

In that sense, we define a synthetic dataset where the inputs $x \in \mathbb{R}^5$ and labels $y \in \mathbb{R}$ are sampled from uniform distributions. As a result, this dataset does not contain any symmetry in its map from input to output, and, ideally, a symmetry discovery method should learn a basis that corresponds to the 0 matrix, so that when sampling group elements, an identity transformation is applied, as no other similarly parameterized transformation should be strictly beneficial for the downstream performance. Nevertheless, SEMoLA’s implementation enforces a constant norm on the Lie algebra basis during training, and, thus, that cannot occur. What we observe instead is that in this case the norm of the matrix concentrates in a single entry, which changes in every one of the runs with different random seeds that we executed, as we can see in Figure 19.

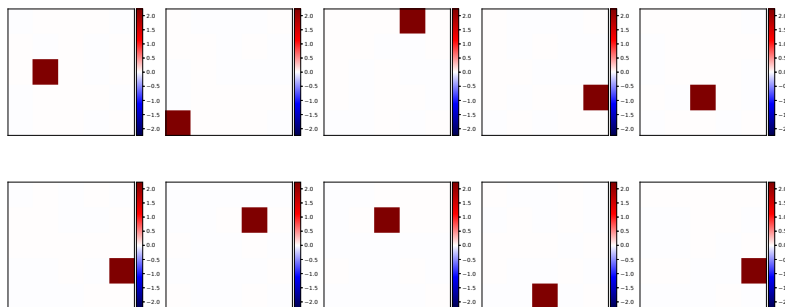


Figure 19: Lie algebras learned in each of 10 runs with different random seeds for the synthetic dataset without symmetries.

As a result, these observations hint that our method could potentially also be applied as a way of determining the extent of symmetry in a dataset, i.e., if the learned Lie algebra basis is similar across runs, the data potentially has a stronger symmetry with respect to the prediction task.

A.11 QUANTUM FIELD THEORY AMPLITUDE PREDICTION

As an additional experiment, we study the problem of defining a neural surrogate for quantum field theoretical amplitudes (Spinner et al., 2024; Aylett-Bullock et al., 2021; Badger and Bullock, 2020), which is of great importance in the physics community due to it being central to the theoretical predictions that LHC measurements are compared to. In particular, these amplitudes describe the un-normalized probability of interactions of fundamental particles as a function of their four-momenta, which are exactly Lorentz-invariant. Following Spinner et al. (2024), we study the setting $q\bar{q} \rightarrow Z + ng$, i.e., the production of a Z boson with $n = 1, \dots, 4$ additional gluons from a quark-antiquark pair. In particular, we focus on the scenario where $n = 2$ and we generate and split the dataset as in that work.

We train SEMoLA using the baseline Transformer model defined in Spinner et al. (2024) as the base prediction model, and with a learnable Lie algebra basis of cardinality $C = 6$ corresponding to that of the restricted Lorentz group $SO^+(1, 3)$, i.e., the connected identity component of the Lorentz group. From Figure 20, we can observe that SEMoLA is able to approximately learn the correct group. Even though some of the elements of the learned Lie algebra basis (especially those corresponding to boosts) are not exactly equivalent to their ground truth counterparts, we believe that this is more of a side-effect caused by the symmetries that exist in this particular dataset rather than a partial imprecision of our method.

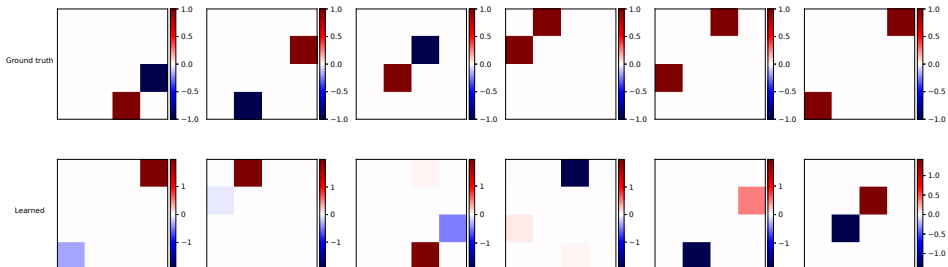


Figure 20: Comparison of the ground truth Lie algebra basis with the one learned by SEMoLA in the QFT amplitudes dataset.

This hypothesis is partly supported by the results in terms of the test error comparison with other baseline models (L-GATr (Spinner et al., 2024) and Transformers with and without data augmentation) that we include in Figure 21, which show that a Transformer trained with augmentations from the ground truth restricted Lorentz group (Transformer_aug) performs worse than the same Transformer without data augmentation. As a result, it is possible that the characteristics of this dataset are such that the samples do not necessarily reflect the action of the entire group, which could potentially explain both the higher accuracy of certain elements of the Lie algebra basis learned by SEMoLA, as well as the slightly negative effect of data augmentation.

1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889

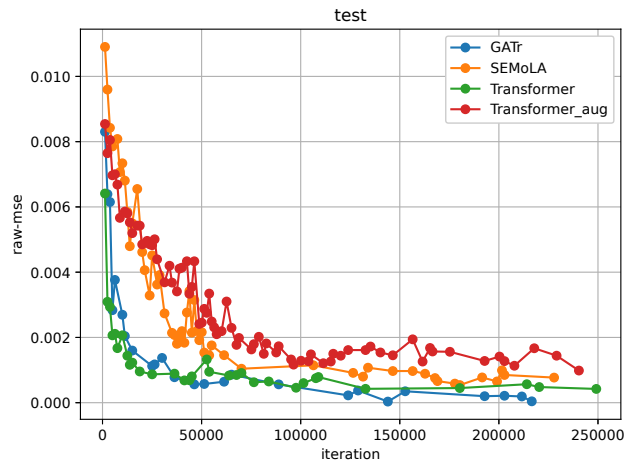


Figure 21: Comparison of the test MSE for different baseline models and SEMoLA in the QFT amplitudes dataset.