Data Augmentation and Regularization for Learning Group Equivariance

Oskar Nordenfors, Axel Flinth
Umeå University, Department of Mathematics and Mathematical Statistics

***Abstract*—In many machine learning tasks, known symmetries can be used as an inductive bias to improve model performance. In this paper, we consider learning group equivariance through training with data augmentation. We summarize results from a previous paper of our own, and extend the results to show that equivariance of the trained model can be achieved through training on augmented data in tandem with regularization.**

I. Introduction

In certain machine learning tasks, symmetries of the data distribution are known a priori. For example, when estimating energy levels of a molecule, they should not change through a rotation. For such tasks, *group equivariant models* have achieved state-of-the-art performance. A prominent example is the AlphaFold 2 model [1]. Group equivariant models have been the focus of much research, and there are many ways to make models equivariant.

One line of research, starting with [2], [3], considers restricting the linear layers of neural networks to ensure that they are equivariant regardless of the values of the learned weights. That is, to achieve equivariance in the model, one makes the model equivariant by design, as opposed to, say, learning the equivariance. One benefit of equivariance-by-design is that there is a clear guarantee that the model will be exactly equivariant throughout training.

Another way to approach the problem is to consider learning equivariance from data. Intuitively, one way to do this is to use data augmentation. This entails supplementing the data through transformations of the inputs and outputs in order to symmetrize the training data distribution. Data augmentation was put into a group-theoretic context in [4], which has led to a great deal of research. Several works have since dealt with the question whether data augmentation (provably) induces equivariance. However, previous studies have been confined to simpler settings, such as linear models [5], [6] as well as linear neural networks [7], [8], that is neural networks without activation functions. However, results for bona fide neural networks are scarce. Despite this, data augmentation is widely used and to great effect, for example in the recent AlphaFold 3 model [9]. This motivates searching for theoretical results also for 'real' neural networks.

In [10], the authors of this paper developed a framework for studying the effects of data augmentation also for more general neural networks. The main results were that, under some geometrical conditions on the nominal neural network architecture, (i) the set of equivariant architectures $\mathcal{E}$ is an invariant subset for the dynamics of gradient descent with augmented data and (ii) the set of stationary points on $\mathcal{E}$ are the same for augmented and equivariant dynamics. $\mathcal{E}$ could however be unstable for the augmented dynamics.

In this paper, we will present this framework and these results of [10], and also study the effect of a simple regularization procedure. We will show that using this strategy in tandem with data augmentation will make the set of equivariant architectures $\mathcal{E}$ for attractor for the training dynamics. We will also provide a small numerical experiment confirming our results.

Although we will not go into this further in this article, let us also mention another recent development related to so-called *ensembles*. In that setting, it is possible to show group equivariance through data augmentation [11]–[13] in a global sense. These results rely heavily on taking averages over many individually trained networks, which can be expensive. This paper treats instead individual networks.

II. Background

We will here present the framework and main results of [10]. Let us begin by defining what we mean by group representation and group equivariance. These are standard definitions and can be found in, for example, [14, p. 3].

Definition 1. A *representation* (*rep*) of a group G on a vector space V is a group homomorphism $\rho : G \to \mathrm{Aut}(V)$. That is, a map that associates each group element with an invertible matrix, in a way that respects the group structure.

If G is compact, the reps can be assumed to be unitary w.l.o.g. In the sequel, we will assume that G is compact. There is always a *trivial rep* defined by $\rho(g) = \mathrm{id}$ for every $g \in G$.

Definition 2. Given a group G and vector spaces U and V with reps ρ_U and ρ_V respectively, a map $f : U \to V$ is called *equivariant* if

$$f(\rho_U(g)u) = \rho_V(g)f(u), \qquad g \in G \,\text{and}\, u \in U, \tag{1}$$

If in (1) ρ_V is the trivial rep, then f is called *invariant*. Given reps ρ_U and ρ_V on U and V, respectively, we define a rep on the space of linear maps $U \to V$ by $\overline{\rho}(g)A = \rho_V(g)^{-1}A\rho_U(g)$. The subspace of equivariant linear maps can then be identified as $\mathrm{Hom}_G(U,V) \coloneqq \{A \in \mathrm{Hom}(U,V) : \overline{\rho}(g)A = A \ \forall g \in G\}$.

A. A neural network framework

Let $X_0, \dots, X_L$ be the input, intermediate, and output spaces of the network, respectively, and $\sigma_i : X_{i+1} \to X_{i+1}$,

$i \in [L]$ be its non-linearities. Denoting the learnable linear layers of the network $A_i : X_i \to X_{i+1}$, we obtain a neural network by recursively defining $x_0 = x$, $x_{i+1} = \sigma_{i+1}(A_i x_i)$, and finally $\Phi(x) = x_L$. Given reps ρ_0 and ρ_L of the group G on the input and output spaces, respectively, it is straightforward to see that if we specify a rep ρ_i of G on all intermediate spaces and choose both the non-linearities σ_i and the linearities A_i are chosen equivariant, Ψ_A will also be. This is essentially the canonical way to construct manifestly equivariant networks [15, p. 27]. Here we refer to the network being trained in *equivariant* mode. We will compare this to the corresponding *nominal* architectures, that is, networks with the same non-linearities, but linear layers not necessarily restricted to the space $\mathcal{H}_G$ of (tuples of) equivariant linear maps.

While the above construction is general, it only entails fully-connected networks without bias. In [10], we showed that a vast range of architectures can painlessly be incorporated into the framework by a priori restraining the linear layers A_i to lie in an affine subspace $\mathcal{L}$. As a simple example, we obtain CNN:s by letting $\mathcal{L}$ be the subspace of convolution operators. We put these into equivariant mode via constraining the layers to lie in the affine subspace $\mathcal{E} = \mathcal{H}_G \cap \mathcal{L}$.

B. Training dynamics

Let $\ell : X_L \times X_L \to \mathbb{R}$ be a loss function. Using training data and labels (x, y) distributed according to a distribution $\mathcal{D}$, we define the *nominal risk* $R(A) \coloneqq \mathbb{E}_{\mathcal{D}}[\ell(\Phi_A(x_0), x_L)]$. Data augmentation is performed by applying symmetry transformations $\rho_0(g), \rho_L(g)$, drawn according to a measure μ. This transforms the risk function into

$$R^{\mathrm{aug}}(A) \coloneqq \mathbb{E}_{\mu}[\mathbb{E}_{\mathcal{D}}[\ell(\Phi_A(\rho_0(g)x_0), \rho_L(g)x_L)]],$$

that we will refer to as the *augmented risk*. We will always assume that μ is the so-called *Haar measure* of the group, which has the property that group translations $h \mapsto gh$ are measure preserving. This is a probability measure, since we have assumed that G is compact. If G is finite, the Haar measure is simply the uniform measure.

Now, let us consider three approaches for training our network. Firstly, we can train our network on non-augmented data via applying gradient flow with the nominal risk. To confine the flow to the subspace $\mathcal{L}$, this gradient flow needs to be projected: Letting $\Pi_{\mathcal{L}}$ denote the orthogonal projection onto $T\mathcal{L}$ (the tangent space of the linear manifold $\mathcal{L}$), we hence consider $\dot{A} = -\Pi_{\mathcal{L}} \nabla R(A)$. In the same manner, training on augmented data amounts to the dynamics $\dot{A} = -\Pi_{\mathcal{L}} \nabla R^{\mathrm{aug}}(A)$. Lastly, we can train our network in equivariant mode via applying gradient flow projected to $\mathcal{E}$, that is $\dot{A} = -\Pi_{\mathcal{E}} \nabla R(A)$, with $\Pi_{\mathcal{E}}$ the orthogonal projection onto $\mathrm{T}\mathcal{E}$. These projections do not need to be applied explicitly – as we discuss in [10], this behavior emerges simply by orthogonal parametrizations of the layers and applying gradient descent to the coefficients of those parametrizations [10, Sec. 3].

C. The 'local equivalence' of augmenting and restricting.

The analysis we performed in [10] revolves around the augmented and restricted dynamics on $\mathcal{E}$. In essence, we proved three results. To arrive at them, we must assume that

(a) The loss function ℓ is invariant:

$$\ell(\rho_L(g)x, \rho_L(g)x') = \ell(x, x'), g \in G, x, x' \in X_L.$$

(b) The nominal architecture satisfies the *compatibility condition*, that is *the orthogonal projections $\Pi_{\mathcal{L}}$ onto $T\mathcal{L}$ and Π_G onto $\mathcal{H}_G$ should commute*.

$$\Pi_{\mathcal{L}} \Pi_G = \Pi_G \Pi_{\mathcal{L}}.$$

This is satisfied if all symmetry transformations $\rho(g)$ leave $\mathcal{L}$ invariant. This shows that the results are applicable to, for example, convolutional neural networks and rotations in image space, as long as the supports of the convolutions are rotationally symmetric.

Under these assumptions, we have the following.

Theorem 1. *1. $\mathcal{E}$ is an invariant set of the augmented dynamics.*

2. The set of points S^{eq} and S^{aug} in $\mathcal{E}$ that are stationary for the equivariant and augmented dynamics, respectively, agree.

Both of these results follow from the (non-trivial) fact, which we will use in the following:

Fact A $\Pi_{\mathcal{L}} \nabla R^{\mathrm{aug}}(A) = \Pi_{\mathcal{E}} \nabla R(A)$ for $A \in \mathcal{E}$.

Note that the above even shows that when restricted to $\mathcal{E}$, the dynamics of the augmented and equivariant training are exactly the same. However, the result only concerns stationarity, which is, of course, different from stability. [10] contains a weak result about stability.

Theorem 2. *Points on $\mathcal{E}$ that are stable under the augmented dynamics are also stable under the equivariant dynamics, but not necessarily the other way around.*

In particular, $\mathcal{E}$ is not guaranteed to be an attractor for the augmented dynamics. In the next section, we show that we can achieve that through regularization.

III. On the Effects of Regularization

A standard way of regularizing neural network training is weight decay, which refers to adding a regularization term proportional to the squared norm of the weights, $\gamma/2 \|A\|^2$ to the risk. This prevents the weights from getting to large. Here, we are interested in the *non-equivariant* parts of the weight getting too large, which suggests using a term of the form $\gamma/2 \|\Pi_{\mathcal{E}^\perp} A\|^2$.

It should be noted that it is not hard to calculate the projection onto $\mathcal{E}^\perp$ explicitly. For most important symmetry groups and architectures, orthogonal bases of $\mathcal{E}$ are known, see for example [16], [17]. There are even methods to calculate them numerically [18].

Let us now give a formal proof that given a strong enough regularization, $\mathcal{E}$ will become an attractor. We will make the following simplifying assumptions.

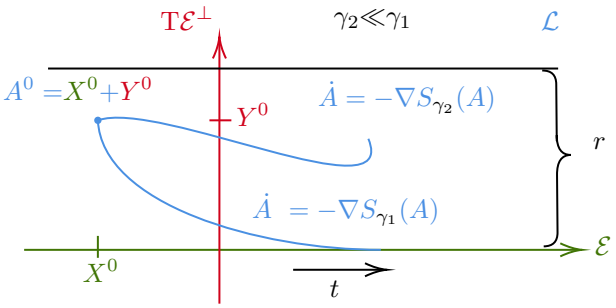

Fig. 1. Regardless of how far from $\mathcal{E}$ we start training we can select the regularization parameter γ large enough that the dynamics $\dot{A} = -\nabla S_\gamma(A)$ converge to $\mathcal{E}$ exponentially fast. If a too low value of γ is chosen, the dynamics might not converge to $\mathcal{E}$ at all.

(i) The risk is bounded below by zero.

(ii) The second derivative of the augmented risk is bounded below in the following sense: There exists a (*not necessarily positive*) constant σ such that

$$\langle (R^{\text{aug}})''(A)Y, Y\rangle \geq \sigma \|Y\|^2, \quad A \in \mathcal{E}, Y \in T\mathcal{E}^\perp.$$

(iii) The third derivative of the augmented risk is uniformly bounded.

Before we come to the main result, let us present another fact derived in [10, proof of Prop. 3.14] that will be used in the proof.

Fact B If $Y \in T\mathcal{E}^\perp$ and $A \in \mathcal{E}$, $\Pi_{\mathcal{L}}(R^{\text{aug}})''(A)Y \in T\mathcal{E}^\perp$. Let us also introduce the notation $\Pi_{\mathcal{E}^\perp}$ for the orthogonal projection onto $T\mathcal{E}^\perp$.

We can now prove the main result, that the equivariant subspace $\mathcal{E}$ becomes an attractor for training with data augmentation by penalizing non-equivariance through a regularization term in the augmented loss.

Theorem 3 (Equivariant subspace is attractor of regularized augmented gradient flow)**.** *Suppose that $R^{\text{aug}}(A) \geq 0$ for every $A \in \mathcal{L}$ and let $S_\gamma(A) = S(A) := R^{\text{aug}}(A) + \frac{\gamma}{2}\|\Pi_{\mathcal{E}^\perp} A\|^2$ denote the regularized loss. Assume that ℓ is invariant, that the compatibility condition $\Pi_{\mathcal{L}}\Pi_G = \Pi_G\Pi_{\mathcal{L}}$ holds, and (i)-(iii) above. Then, for any $r > 0$, we can choose $\gamma > 0$ large enough so that any curve started at a distance to $\mathcal{E}$ smaller than r following the dynamics*

$$\dot{A} = -\nabla S(A) = -\Pi_{\mathcal{L}}\nabla R^{\text{aug}}(A) - \gamma \Pi_{\mathcal{E}^\perp} A$$

will converge to $\mathcal{E}$ exponentially fast.

Proof. Let $X \in \mathcal{E}$ and $Y \in T\mathcal{E}^\perp$ and write $A = X + Y$, so that $S(A) = R^{\text{aug}}(A) + \frac{\gamma}{2}\|Y\|^2$. Due to (i), we then have

$$\frac{\gamma}{2}\|Y\|^2 \leq R^{\text{aug}}(A) + \frac{\gamma}{2}\|Y\|^2 \leq R^{\text{aug}}(A^0) + \frac{\gamma}{2}\|Y^0\|^2,$$

where $A^0 = X^0 + Y^0$ is the starting point of the gradient flow $\dot{A} = -\nabla S(A)$. Thus, we get the a priori bound

$$\|Y\|^2 \leq \frac{2}{\gamma} R^{\text{aug}}(A^0) + \|Y^0\|^2 =: \alpha$$

for every A. Now, a Taylor expansion of $\nabla R^{\text{aug}}(A)$ around X yields

$$\nabla R^{\text{aug}}(A) = \Pi_{\mathcal{L}}\nabla R^{\text{aug}}(X) + \Pi_{\mathcal{L}}(R^{\text{aug}})''(X)Y + O(\|Y\|^2).$$

Note that the big-O is independent of X due to assumption (iii). Due to Facts A and B above, we see that $\Pi_{\mathcal{L}}\nabla R^{\text{aug}}(X) = \Pi_{\mathcal{E}}\nabla R(A) \in T\mathcal{E}$ and $\Pi_{\mathcal{L}}(R^{\text{aug}})''(X)Y \in T\mathcal{E}^\perp$. Hence, as noted in to [10, Prop. 3.14], the dynamics $\dot{A} = -\nabla S(A)$ decouple in the following sense:

$$\begin{aligned} \dot{X} &= -\Pi_{\mathcal{E}}\nabla R(A) + O(\|Y\|^2), \\ \dot{Y} &= -\Pi_{\mathcal{L}}(R^{\text{aug}})''(X)Y - \gamma Y + O(\|Y\|^2). \end{aligned} \tag{2}$$

Note that the term $-\gamma Y$ is due to the regularization. Differentiating the squared norm of Y, (2) yields

$$\begin{aligned} &\frac{d}{dt}\Big(\frac{1}{2}\|Y\|^2\Big) = \langle \dot{Y}, Y\rangle \\ &= -\langle \Pi_{\mathcal{L}}(R^{\text{aug}})''(X)Y, Y\rangle - \langle \gamma Y, Y\rangle + \langle O(\|Y\|^2), Y\rangle \\ &\leq -\sigma\|Y\|^2 - \gamma\|Y\|^2 + C\sqrt{\alpha}\|Y\|^2 \\ &= (C\sqrt{\alpha} - \sigma - \gamma)\|Y\|^2, \end{aligned} \tag{3}$$

where we used the a priori bound on $\|Y\|$ and assumption (ii). From (3) it follows by Grönwall's inequality that

$$\|Y\|^2 \leq \|Y^0\|^2 \exp\Big(2(C\sqrt{\alpha} - \sigma - \gamma)t\Big).$$

Hence, if $\gamma > C\sqrt{\alpha} - \sigma$, $\|Y\|^2$ decays exponentially to 0. Since α only depends on $\|Y^0\|$, we see that we can achieve that by choosing γ in dependence of r, which is what was to be proved. □

Remark 1. From an intuitive point of view, heavily regularizing the non-equivariant part $\Pi_{\mathcal{E}^\perp} A$ of the linear layers will also force the dynamics down to $\mathcal{E}$ in the non-augmented case. This is to some degree true, but the augmented case is different from the non-augmented one in two crucial ways. To see this, let us, in the same way as in the proof of Theorem 3 Taylor expand the gradient of the nominal risk $\nabla R(A)$ around an $X \in \mathcal{E}$ and write down the Y dynamics as

$$\dot{Y} = -\Pi_{\mathcal{L}}\nabla R(X) - \Pi_{\mathcal{L}} R''(X)Y - \gamma Y + O(\|Y\|^2),$$

which yields

$$\begin{aligned} \frac{d}{dt}\Big(\frac{1}{2}\|Y\|^2\Big) = &-\langle \Pi_{\mathcal{L}}\nabla R(X), Y\rangle - \langle \Pi_{\mathcal{L}} R''(X)Y, Y\rangle \\ &- \langle \gamma Y, Y\rangle + \langle O(\|Y\|^2), Y\rangle. \end{aligned}$$

Now, in the non-augmented case we are not guaranteed that $\langle \mathcal{L}\nabla R(X), Y\rangle = 0$, as opposed to the augmented case where Fact A ensures that the inner product vanishes.

Assuming $\langle \Pi_{\mathcal{L}}\nabla R(X), Y\rangle = 0$, an argument similar to that in the proof of Theorem 3 would yield a similar result. However, in this case, the parameter σ would be different and would be a smaller number (which would lead to a higher required value of γ). To see this, let us use the following formula, which follows by [10, Lemma 3.11]

$$\langle (R^{\text{aug}})''(A)Y, Y\rangle = \int_G \langle R''(A)\rho(g)Y, \rho(g)Y\rangle d\mu(g)$$

Now, if $Y \in T\mathcal{E}^\perp$, it is not hard to see that $\rho(g)Y \in T\mathcal{E}^\perp$ also for all g. Hence, if $\langle R''(A)Y, Y\rangle \geq \overline{\sigma}\|Y\|^2$ for all $Y \in T\mathcal{E}^\perp, A \in \mathcal{E}$, the above integral is bounded below by $\overline{\sigma}\|Y\|^2$. In this sense, $\sigma \geq \overline{\sigma}$, and therefore we can potentially get away with a lower value for γ when regularizing the augmented risk compared to the nominal risk.

Remark 2. We can also say something about what happens if we begin training in a neighborhood in $\mathcal{L}$ of a strict local minimum X^* for the equivariant mode training. Let us write $A = (X, Y)$, where $X \in \mathcal{E}$ and $Y \in \mathrm{T}\mathcal{E}^\perp$. Consider the dynamics

$$\dot{X} = -\Pi_{\mathcal{E}} \nabla S(X, Y) = -\Pi_{\mathcal{E}} \Pi_{\mathcal{L}} \nabla R^{\mathrm{aug}}(X, Y),$$
$$\dot{Y} = -\Pi_{\mathcal{E}^\perp} \nabla S(X, Y) = -\Pi_{\mathcal{E}^\perp} \Pi_{\mathcal{L}} \nabla R^{\mathrm{aug}}(X, Y) - \gamma Y.$$

Let $\Delta X = X - X^*$. Taylor expanding around $(X^*, 0)$ gives

$$\begin{bmatrix} \dot{\Delta X} \\ \dot{Y} \end{bmatrix} = -\left((R^{\mathrm{aug}})''(X^*, 0) + \begin{bmatrix} 0 & 0 \\ 0 & \gamma I \end{bmatrix} \right) \begin{bmatrix} \Delta X \\ Y \end{bmatrix} + O(\|\Delta X\|^2 + \|Y\|^2)$$

By choosing γ large enough, we can ensure that the second derivative is positive definite. Thus, there is a neighborhood U of the point $(X^*, 0)$ such that starting the training at a point $(X^0, Y^0) \in U$, we guarantee that the regularized augmented training will converge to $(X^*, 0)$.

IV. Experimenting with Different Values of γ

We perform a numerical experiment to confirm the practical relevance of our result. In the experiment we use SGD with random augmentations, which is different from what we do in our theoretical development, where we assume training with perfect augmentation and gradient flow, but also closer to what is done in practice. Since our results here in particular are about stability, they should also have an impact in the stochastic setting.

a) Experiment description: We consider CNN:s with two convolutional layers of 16 channels each, with filters of size 3×3, followed by a fully connected layer (both without biases). We use $\tanh$ activation functions and a cross-entropy loss. Since the activation function acts pixel-wise, they are equivariant. The networks are trained in three configurations: (i) in equivariant mode, (ii) on (randomly) augmented data (iii) on non-augmented data. The group action we consider is the discrete group of 90 degree rotations, and aim to build invariant networks. As noted above, these networks obey the compatibility condition. Since the group acts trivially on the output space, the invariance of the loss function ℓ is trivial.

We initialize the equivariant networks with Gaussian weights. We then copy those weights to the non-equivariant networks, and subsequently perturb them with Gaussian noise (independently from each other). We train the networks on augmented and non-augmented data, respectively, using SGD on MNIST for 10 epochs with a learning rate of 1e-3 and a batch size of 10. Both non-equivariant networks are regularized, using four different values of the regularization constant γ, 1e-4,1e-2,1e0 and 1e2. We record after each gradient step the distance of the non-equivariant models to $\mathcal{E}$. The experiment is repeated 30 times for each value of γ. All code used in the experiment is made available at https://github.com/usinedepain/eq_aug_reg_release

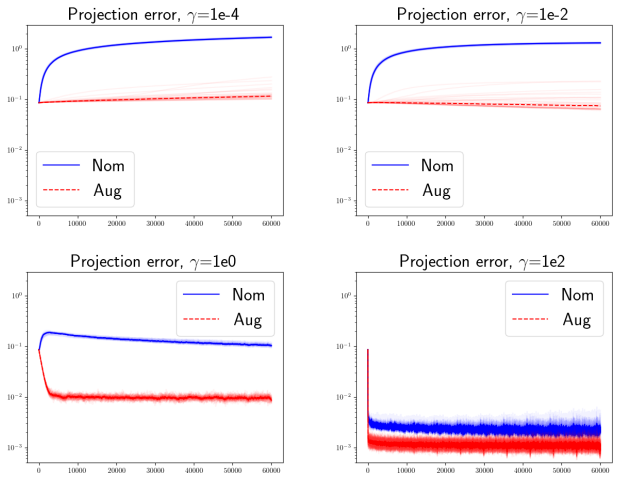

Fig. 2. Projection errors for the two non-equivariant models for different values of γ. Notice the logarithmic y-scale. Opaque lines are medians, and transparent lines are individual runs. Best viewed in color.

b) Results: In Figure 2, we plot the evolution of the distance to $\mathcal{E}$ along the training – the opaque lines are the median values. The graphs look just as one expects – for high values of γ, both non-equivariant models stay close to $\mathcal{E}$. However, already for $\gamma = 1\mathrm{e}0$, the model trained on non-augmented data drifts considerably, while the augmented model stays close to $\mathcal{E}$. We also see that when the constant is chosen too low ($\gamma =$ 1e-4), the augmented model will also drift from $\mathcal{E}$.

V. Conclusion

We investigated the relationship between manifestly equivariant neural network architectures with data augmentation. Using the 'softer' approach of data augmentation and a regularization term to punish distance to an equivariant subspace $\mathcal{E}$, we obtain equivariant models without having to restrict the layers of the network to be equivariant a priori. The theoretical results rely heavily on our previous paper [10]. We have also seen, with a small numerical experiment, that these results are born out in practice, even when using SGD with random augmentations.

For future work, it would be interesting to do larger experiments to see if these methods can give lower test error than manifest equivariance. In particular, it would be interesting to see whether a fine-tuning of γ could make the networks avoid bad local minima on $\mathcal{E}$ while still being attracted to good ones.

VI. Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

REFERENCES

[1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with alphafold," *nature*, vol. 596, pp. 583–589, 2021.

[2] T. S. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proceedings of the 33rd International Conference on Machine Learning*, 2016, pp. 2990–2999.

[3] ——, "Steerable cnns," in *Proceedings of the 5th International Conference on Learning Representations*, 2017, pp. 689–703.

[4] S. Chen, E. Dobriban, and J. H. Lee, "A group-theoretic framework for data augmentation," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 9885–9955, 2020.

[5] C. Lyle, M. van der Wilk, M. Kwiatkowska, Y. Gal, and B. Bloem-Reddy, "On the benefits of invariance in neural networks," *arXiv:2005.00178*, 2020.

[6] B. Elesedy and S. Zaidi, "Provably strict generalisation benefit for equivariant models," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pp. 2959–2969.

[7] H. Lawrence, K. Georgiev, A. Dienes, and B. T. Kiani, "Implicit bias of linear equivariant networks," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 096–12 125.

[8] Z. Chen and W. Zhu, "On the implicit bias of linear equivariant steerable networks," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[9] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick *et al.*, "Accurate structure prediction of biomolecular interactions with alphafold 3," *Nature*, vol. 630, pp. 493–500, 2024.

[10] O. Nordenfors, F. Ohlsson, and A. Flinth, "Optimization dynamics of equivariant and augmented neural networks," *Transactions of Machine Learning Research*, vol. 4, 2025. [Online]. Available: https://arxiv.org/abs/2303.13458

[11] J. E. Gerken and P. Kessel, "Emergent equivariance in deep ensembles," in *Proceedings of the 41st International Conference on Machine Learning*, vol. 235, 2024, pp. 15 438–15 465.

[12] J. Maass and J. Fontbona, "Symmetries in overparametrized neural networks: A mean-field view," 2024. [Online]. Available: https://arxiv.org/abs/2405.19995

[13] O. Nordenfors and A. Flinth, "Ensembles provably learn equivariance through data augmentation," 2024. [Online]. Available: https://arxiv.org/abs/2410.01452

[14] W. Fulton and J. Harris, "Representation theory," *Graduate Texts in Mathematics*, vol. 129, 2004.

[15] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *arXiv preprint arXiv:2104.13478*, 2021.

[16] H. Maron, H. Ben-Hamu, N. Shamir, and Y. Lipman, "Invariant and equivariant graph networks," in *Proceedings of the 7th International Conference on Learning Representations*, 2019.

[17] T. Cohen, M. Geiger, and M. Weiler, "A general theory of equivariant CNNs on homogeneous spaces," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[18] M. Finzi, M. Welling, and A. G. Wilson, "A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pp. 3318–3328.