EFFICIENT GENERALIZED SPHERICAL CNNS

Oliver J. Cobb, Christopher G. R. Wallis, Augustine N. Mavor-Parker,
Augustin Marignier, Matthew A. Price, Mayeul d'Avezac & Jason D. McEwen*
Kagenova Limited, Guildford GU5 9LD, UK

ABSTRACT

Many problems across computer vision and the natural sciences require the analysis of spherical data, for which representations may be learned efficiently by encoding equivariance to rotational symmetries. We present a generalized spherical CNN framework that encompasses various existing approaches and allows them to be leveraged alongside each other. The only existing non-linear spherical CNN layer that is strictly equivariant has complexity $\mathcal{O}(C^2L^5)$, where C is a measure of representational capacity and L the spherical harmonic bandlimit. Such a high computational cost often prohibits the use of strictly equivariant spherical CNNs. We develop two new strictly equivariant layers with reduced complexity $\mathcal{O}(CL^4)$ and $\mathcal{O}(CL^3 \log L)$, making larger, more expressive models computationally feasible. Moreover, we adopt efficient sampling theory to achieve further computational savings. We show that these developments allow the construction of more expressive hybrid models that achieve state-of-the-art accuracy and parameter efficiency on spherical benchmark problems.

1 INTRODUCTION

Many fields involve data that live inherently on spherical manifolds, e.g. 360° photo and video content in virtual reality and computer vision, the cosmic microwave background radiation from the Big Bang in cosmology, topographic and gravitational maps in planetary sciences, and molecular shape orientations in molecular chemistry, to name just a few. Convolutional neural networks (CNNs) have been tremendously effective for data defined on Euclidean domains, such as the 1D line, 2D plane, or nD volumes, thanks in part to their translation invariance properties. However, these techniques are not effective for data defined on spherical manifolds, which have a very different geometric structure to Euclidean spaces (see Appendix A). To transfer the remarkable success of deep learning to data defined on spherical domains, deep learning techniques defined inherently on the sphere are required. Recently, a number of spherical CNN constructions have been proposed.

Existing CNN constructions on the sphere fall broadly into three categories: fully real (i.e. pixel) space approaches (e.g. Boomsma & Frellsen, 2017; Jiang et al., 2019; Perraudin et al., 2019; Cohen et al., 2019); combined real and harmonic space approaches (Cohen et al., 2018; Esteves et al., 2018; 2020); and fully harmonic space approaches (Kondor et al., 2018). Real space approaches can often be computed efficiently but they necessarily provide an approximate representation of spherical signals and the connection to the underlying continuous symmetries of the sphere is lost. Consequently, such approaches cannot fully capture rotational equivariance. Other constructions take a combined real and harmonic space approach (Cohen et al., 2018; Esteves et al., 2018; 2020), where sampling theorems (Driscoll & Healy, 1994; Kostelec & Rockmore, 2008) are exploited to connect with underlying continuous signal representations to capture the continuous symmetries of the sphere. However, in these approaches non-linear activation functions are computed pointwise in real space, which induces aliasing errors that break strict rotational equivariance. Fully harmonic space spherical CNNs have been constructed by Kondor et al. (2018). A continual connection with underlying continuous signal representations is captured by using harmonic signal representations throughout. Consequently, this is the only approach exhibiting strict rotational equivariance. However, strict equivariance comes at great computational cost, which can often prohibit usage.

*Corresponding author: jason.mcewen@kagenova.com

In this article we present a generalized framework for CNNs on the sphere (and rotation group), which encompasses and builds on the influential approaches of Cohen et al. (2018), Esteves et al. (2018) and Kondor et al. (2018) and allows them to be leveraged alongside each other. We adopt a harmonic signal representation in order to retain the connection with underlying continuous representations and thus capture all symmetries and geometric properties of the sphere. We construct new fully harmonic (non-linear) spherical layers that are strictly rotationally equivariant, are parameter-efficient, and dramatically reduce computational cost compared to similar approaches. This is achieved by a channel-wise structure, constrained generalized convolutions, and an optimized degree mixing set determined by a minimum spanning tree. Furthermore, we adopt efficient sampling theorems on the sphere (McEwen & Wiaux, 2011) and rotation group (McEwen et al., 2015a) to improve efficiency compared to the sampling theorems used in existing approaches (Driscoll & Healy, 1994; Kostelec & Rockmore, 2008). We demonstrate state-of-the-art performance on all spherical benchmark problems considered, both in terms of accuracy and parameter efficiency.

2 Generalized Spherical CNNs

We first overview the theoretical underpinnings of the spherical CNN frameworks introduced by Cohen et al. (2018), Esteves et al. (2018), and Kondor et al. (2018), which make a connection to underlying continuous signals through harmonic representations. For more in-depth treatments of the underlying harmonic analysis we recommend Esteves (2020), Kennedy & Sadeghi (2013) and Gallier & Quaintance (2019). We then present a generalized spherical layer in which these and other existing frameworks are encompassed, allowing existing frameworks to be easily integrated and leveraged alongside each other in hybrid networks.

Throughout the following we consider a network composed of S rotationally equivariant layers $\mathcal{A}^{(1)}, ..., \mathcal{A}^{(S)}$, where the i-th layer $\mathcal{A}^{(i)}$ maps an input activation $f^{(i-1)} \in \boldsymbol{\mathcal{H}}^{(i-1)}$ onto an output activation $f^{(i)} \in \boldsymbol{\mathcal{H}}^{(i)}$. We focus on the case where the network input space $\boldsymbol{\mathcal{H}}^{(0)}$ consists of spherical signals (but note that input signals on the rotation group may also be considered).

2.1 Signals on the Sphere and Rotation Group

Let $L^2(\Omega)$ denote the space of square-integrable functions over domain Ω. A signal $f \in L^2(\Omega)$ on the sphere ($\Omega = \mathbb{S}^2$) or rotation group ($\Omega = \mathrm{SO}(3)$) can be rotated by $\rho \in \mathrm{SO}(3)$ by defining the action of rotation on signals by $\mathcal{R}_\rho f(\omega) = f(\rho^{-1}\omega)$ for $\omega \in \Omega$. An operator $\mathcal{A} : L^2(\Omega_1) \to L^2(\Omega_2)$, where $\Omega_1, \Omega_2 \in \{\mathbb{S}^2, \mathrm{SO}(3)\}$, is then equivariant to rotations if $\mathcal{R}_\rho(\mathcal{A}(f)) = \mathcal{A}(\mathcal{R}_\rho f)$ for all $f \in L^2(\Omega_1)$ and $\rho \in \mathrm{SO}(3)$, i.e. rotating the function before application of the operator is equivalent to application of the operator first, followed by a rotation.

A spherical signal $f \in L^2(\mathbb{S}^2)$ admits a harmonic representation $(\hat{f}^0, \hat{f}^1, ...,)$ where $\hat{f}^\ell \in \mathbb{C}^{2\ell+1}$ are the harmonic coefficients given by the inner product $\langle f, Y_m^\ell \rangle$, where Y_m^ℓ are the spherical harmonic functions of degree ℓ and order $|m| \leqslant \ell$. Likewise a signal $f \in L^2(\mathrm{SO}(3))$ on the rotation group admits a harmonic representation $(\hat{f}^0, \hat{f}^1, ...)$ where $\hat{f}^\ell \in \mathbb{C}^{(2\ell+1)\times(2\ell+1)}$ are the harmonic coefficients with (m, n)-th entry $\langle f, D^\ell_{mn} \rangle$ for integers $|m|, |n| \leqslant \ell$, where $D^\ell : \mathrm{SO}(3) \to \mathbb{C}^{(2\ell+1)\times(2\ell+1)}$ is the unique $2\ell + 1$ dimensional irreducible group representation of $\mathrm{SO}(3)$ on $\mathbb{C}^{(2\ell+1)}$. The rotation $f \mapsto \mathcal{R}_\rho f$ of a signal $f \in L^2(\Omega)$ can be described in harmonic space by $\hat{f}^\ell \mapsto D^\ell(\rho)\hat{f}^\ell$.

A signal on the sphere or rotation group is said to be bandlimited at L if, respectively, $\langle f, Y_m^\ell \rangle = 0$ or $\langle f, D^\ell_{mn} \rangle = 0$ for $\ell \geqslant L$. Furthermore, a signal on the rotation group is said to be azimuthally bandlimited at N if, additionally, $\langle f, D^\ell_{mn} \rangle = 0$ for $|n| \geqslant N$. Bandlimited signals therefore admit finite harmonic representations $(\hat{f}^0, ..., \hat{f}^{L-1})$. In practice real-world signals can be accurately represented by suitably bandlimited signals; henceforth, we assume signals are bandlimited.

2.2 Convolution on the Sphere and Rotation Group

A standard definition of convolution between two signals $f, \psi \in L^2(\Omega)$ on either the sphere ($\Omega = \mathbb{S}^2$) or rotation group ($\Omega = \mathrm{SO}(3)$) is given by

$$(f \star \psi)(\rho) = \langle f, \mathcal{R}_\rho \psi \rangle = \int_\Omega \mathrm{d}\mu(\omega) f(\omega)\, \psi^*(\rho^{-1}\omega), \tag{1}$$

where $d\mu(\omega)$ denotes the Haar measure on Ω and $\cdot^*$ complex conjugation (e.g. Wandelt & Górski, 2001; McEwen et al., 2007; 2013; 2015b; 2018; Cohen et al., 2018; Esteves et al., 2018). In particular, the convolution satisfies

$$((\mathcal{R}_\rho f) \star \psi)(\rho') = \langle \mathcal{R}_\rho f, \mathcal{R}_{\rho'} \psi \rangle = \langle f, \mathcal{R}_{\rho^{-1}\rho'} \psi \rangle = (\mathcal{R}_\rho (f \star \psi))(\rho') \tag{2}$$

and is therefore a rotationally equivariant linear operation, which we shall denote by $\mathcal{L}^{(\psi)}$.

The convolution of bandlimited signals can be computed exactly and efficiently in harmonic space as

$$\widehat{(f \star \psi)}^\ell = \hat{f}^\ell \hat{\psi}^{\ell *}, \qquad \ell = 0, ..., L-1, \tag{3}$$

which for each degree ℓ is a vector outer product for signals on the sphere and a matrix product for signals on the rotation group (see Appendix B for further details). Convolving in this manner results in signals on the rotation group (for inputs on both the sphere and rotation group). However, in the spherical case, if the filter is invariant to azimuthal rotations the resultant convolved signal may be interpreted as a signal on the sphere (see Appendix B).

2.3 Generalized Signal Representations

The harmonic representations and convolutions described above have proven useful for describing rotationally equivariant linear operators $\mathcal{L}^{(\psi)}$. Cohen et al. (2018) and Esteves et al. (2018) define spherical CNNs that sequentially apply this operator, with intermediary representations taking the form of signals on $\mathrm{SO}(3)$ and $\mathbb{S}^2$ respectively. Alternatively, for intermediary representations we now consider the more general space of signals introduced by Kondor et al. (2018), to which the aforementioned notions of rotation and convolution naturally extend.

In describing the generalization we first note from Section 2.1 that all bandlimited signals on the sphere and rotation group can be represented as a set of variable length vectors of the form

$$f = \{\hat{f}_t^\ell \in \mathbb{C}^{2\ell+1} : \ell = 0, .., L-1;\ t = 1, ..., \tau_f^\ell\}, \tag{4}$$

where $\tau_f^\ell = 1$ for signals on the sphere and $\tau_f^\ell = \min(2\ell+1, 2N-1)$ for signals on the rotation group. The generalization is to let $\boldsymbol{\mathcal{F}}_L$ be the space of all such sets of variable length vectors, with τ_f unrestricted. This more general space contains the spaces of bandlimited signals on the sphere and rotation group as strict subspaces. For a generalized signal $f \in \boldsymbol{\mathcal{F}}_L$ we adopt the terminology of Kondor et al. (2018) by referring to $\hat{f}_t^\ell$ as the t-th fragment of degree ℓ and to $\tau_f = (\tau_f^0, ..., \tau_f^{L-1})$, specifying the number of fragments for each degree, as the type of f. The action of rotations upon $\boldsymbol{\mathcal{F}}_L$ can be naturally extended from their action upon $L^2(\mathbb{S}^2)$ and $L^2(\mathrm{SO}(3))$. For $f \in \boldsymbol{\mathcal{F}}_L$ we define the rotation operator $f \mapsto \mathcal{R}_\rho f$ by $\hat{f}_t^\ell \mapsto D^\ell(\rho)\hat{f}_t^\ell$, allowing us to extend the usual notion of equivariance to operators $\mathcal{A} : \boldsymbol{\mathcal{F}}_L \to \boldsymbol{\mathcal{F}}_L$.

2.4 Generalized Convolutions

The convolution described by Equation 1 provides a learnable linear operator $\mathcal{L}^{(\psi)}$ that satisfies the desired property of equivariance. Nevertheless, given the generalized interpretation of signals on $\mathbb{S}^2$ and $\mathrm{SO}(3)$ as signals in $\boldsymbol{\mathcal{F}}_L$, the notion of convolution can also be generalized (Kondor et al., 2018).

In order to linearly and equivariantly transform a signal $f \in \boldsymbol{\mathcal{F}}_L$ of type τ_f into a new signal $f * \psi \in \boldsymbol{\mathcal{F}}_L$ of any desired type $\tau_{(f*\psi)}$, we may specify a filter $\psi = \{\hat{\psi}^\ell \in \mathbb{C}^{\tau_f^\ell \times \tau_{(f*\psi)}^\ell} : \ell = 0, ..., L-1\}$, which in general is not an element of $\boldsymbol{\mathcal{F}}_L$, and define a transformation $f \mapsto f * \psi$ by

$$\widehat{(f * \psi)}_t^\ell = \sum_{t'=1}^{\tau_f^\ell} \hat{f}_{t'}^\ell \, \hat{\psi}_{t,t'}^\ell, \qquad \ell = 0, ..., L-1;\ \ t = 1, ..., \tau_{(f*\psi)}^\ell. \tag{5}$$

The degree-ℓ fragments of the transformed signal $(f * \psi)$ are simply linear combinations of the degree-ℓ fragments of f, with no mixing between degrees. Equation 3 shows that this is precisely the form taken by convolution on the sphere and rotation group. In fact Kondor & Trivedi (2018) show that all equivariant linear operations take this general form; the standard notion of convolution is just a special case. One benefit to the generalized notion is that the filter ψ is not forced to occupy the same domain as the signal f, thus allowing control over the type $\tau_{(f*\psi)}$ of the transformed signal. We use $\mathcal{L}_{\mathrm{G}}^{(\psi)}$ to denote this generalized convolutional operator.

2.5 NON-LINEAR ACTIVATION OPERATORS

For $\boldsymbol{\mathcal{F}}_L$ to be a useful representational space, it must be possible to not only linearly but also non-linearly transform its elements in an equivariant manner. However, equivariance and non-linearity is not enough. Equivariant linear operators cannot mix information corresponding to different degrees. Therefore it is of crucial importance that degree mixing is achieved by the non-linear operator.

2.5.1 POINTWISE ACTIVATIONS

When the type τ_f of $f \in \boldsymbol{\mathcal{F}}_L$ permits an interpretation as a signal on $\mathbb{S}^2$ or $\mathrm{SO}(3)$ we may perform an inverse harmonic transform to map the function onto a sample-based representation (e.g. Driscoll & Healy, 1994; McEwen & Wiaux, 2011; Kostelec & Rockmore, 2008; McEwen et al., 2015a). A non-linear function $\sigma : \mathbb{C} \to \mathbb{C}$ may then be applied pointwise, i.e. separately to each sample, before performing a harmonic transform to return to a representation in $\boldsymbol{\mathcal{F}}_L$. We denote the corresponding non-linear operator as $\mathcal{N}_\sigma(f) = \mathcal{F}(\sigma(\mathcal{F}^{-1}(f)))$, where $\mathcal{F}$ represents the harmonic (i.e. Fourier) transform on $\mathbb{S}^2$ or $\mathrm{SO}(3)$. The computational cost of the non-linear operator is dominated by the harmonic transforms. While costly, fast algorithms can be leveraged (see Appendix A). While inverse and forward harmonic transforms on $\mathbb{S}^2$ or $\mathrm{SO}(3)$ that are based on a sampling theory maintain perfect equivariance for bandlimited signals, the pointwise application of σ (most commonly ReLU) is only equivariant in the continuous limit $L \to \infty$. For any finite bandlimit L, aliasing effects are introduced such that equivariance becomes approximate only, as shown by the following experiments.

We consider 100 random rotations $\rho \in \mathrm{SO}(3)$, for each of 100 random signal-filter pairs (f, ψ), and compute the mean equivariance error $d(\mathcal{A}(\mathcal{R}_\rho f), \mathcal{R}_\rho(\mathcal{A}f))$ for operator $\mathcal{A}$, where $d(f, g) = \|f - g\|/\|f\|$ is the relative distance between signals. For convolutions the equivariance error is 4.4×10^{-7} for signals on $\mathbb{S}^2$ and 5.3×10^{-7} for signals on $\mathrm{SO}(3)$ (achieving floating point precision). By comparison the equivariance error for a pointwise ReLU is 0.34 for signals on $\mathbb{S}^2$ and 0.37 for signals on $\mathrm{SO}(3)$. Only approximate equivariance is achieved for the ReLU since the non-linear operation spreads information to higher degrees that are not captured at the original bandlimit, resulting in aliasing. To demonstrate this point we reduce aliasing error by oversampling the real-space signal. When oversampling by $2\times$ or $8\times$ for signals on $\mathrm{SO}(3)$ the equivariance error of the ReLU is reduced to 0.10 and 0.01, respectively. See Appendix D for further experimental details.

Despite the high cost of repeated harmonic transforms and imperfect equivariance, this is nevertheless the approach adopted by Cohen et al. (2018), Esteves et al. (2018) and others, who find empirically that such models maintain a reasonable degree of equivariance.

2.5.2 TENSOR-PRODUCT ACTIVATIONS

In order to define a strictly equivariant non-linear operation that can be applied to a signal $f \in \boldsymbol{\mathcal{F}}_L$ of any type τ_f the decomposability of tensor products between group representations may be leveraged, as first considered by Thomas et al. (2018) in the context of neural networks.

Given two group representations D^{ℓ_1} and D^{ℓ_2} of $\mathrm{SO}(3)$ on $\mathbb{C}^{2\ell_1+1}$ and $\mathbb{C}^{2\ell_2+1}$ respectively, the tensor-product group representation $D^{\ell_1} \otimes D^{\ell_2}$ of $\mathrm{SO}(3)$ on $\mathbb{C}^{2\ell_1+1} \otimes \mathbb{C}^{2\ell_2+1}$ is defined such that $(D^{\ell_1} \otimes D^{\ell_2})(\rho) = D^{\ell_1}(\rho) \otimes D^{\ell_2}(\rho)$ for all $\rho \in \mathrm{SO}(3)$. Decomposing $D^{\ell_1} \otimes D^{\ell_2}$ into a direct sum of irreducible group representations then constitutes finding a change of basis for $\mathbb{C}^{2\ell_1+1} \otimes \mathbb{C}^{2\ell_2+1}$ such that $(D^{\ell_1} \otimes D^{\ell_2})(\rho)$ is block diagonal, where for each ℓ there is a block equal to $D^\ell(\rho)$. The necessary change of basis for $\hat{u}^{\ell_1} \otimes \hat{v}^{\ell_2} \in \mathbb{C}^{2\ell_1+1} \otimes \mathbb{C}^{2\ell_2+1}$ is given by

$$(\hat{u}^{\ell_1} \otimes \hat{v}^{\ell_2})^\ell_m = \sum_{m_1=-\ell_1}^{\ell_1} \sum_{m_2=-\ell_2}^{\ell_2} C^{\ell_1,\ell_2,\ell}_{m_1,m_2,m} \hat{u}^{\ell_1}_{m_1} \hat{v}^{\ell_2}_{m_2}, \tag{6}$$

where $C^{\ell_1,\ell_2,\ell}_{m_1,m_2,m} \in \mathbb{C}$ denote Clebsch-Gordan coefficients whose symmetry properties are such that $(\hat{u}^{\ell_1} \otimes \hat{v}^{\ell_2})^\ell_m$ is non-zero only for $|\ell_1 - \ell_2| \leqslant \ell \leqslant \ell_1 + \ell_2$. The use of Equation 6 arises naturally in quantum mechanics when coupling angular momenta.

This property is useful since if $\hat{f}^{\ell_1} \in \mathbb{C}^{2\ell_1+1}$ and $\hat{f}^{\ell_2} \in \mathbb{C}^{2\ell_2+1}$ are two fragments that are equivariant with respect to rotations of the network input, then a rotation of ρ applied to the network input results

in $\hat{f}^{\ell_1} \otimes \hat{f}^{\ell_2}$ transforming as

$$[\hat{f}^{\ell_1} \otimes \hat{f}^{\ell_2}]^{\ell} \mapsto [(D^{\ell_1}(\rho)\hat{f}^{\ell_1}) \otimes (D^{\ell_2}(\rho)\hat{f}^{\ell_2})]^{\ell} \tag{7}$$

$$= [(D^{\ell_1} \otimes D^{\ell_2})(\rho)(\hat{f}^{\ell_1} \otimes \hat{f}^{\ell_2})]^{\ell} \tag{8}$$

$$= D^{\ell}(\rho)[\hat{f}^{\ell_1} \otimes \hat{f}^{\ell_2}]^{\ell}, \tag{9}$$

where the final equality follows by block diagonality with respect to the chosen basis. Therefore, if fragments $\hat{f}^{\ell_1}$ and $\hat{f}^{\ell_2}$ are equivariant with repsect to rotations of the network input, then so is the fragment $(C^{\ell_1,\ell_2,\ell})^{\top}(\hat{f}^{\ell_1} \otimes \hat{f}^{\ell_2}) \in \mathbb{C}^{2\ell+1}$, where we have written Equation 6 more compactly. We now describe how Kondor et al. (2018) use this fact to define equivariant non-linear transformations of elements in $\boldsymbol{\mathcal{F}}_L$.

A signal $f = \{\hat{f}_t^{\ell} \in \mathbb{C}^{2\ell+1} : \ell = 0, .., L-1;\ t = 1, ..., \tau_f^{\ell}\} \in \boldsymbol{\mathcal{F}}_L$ may be equivariantly and non-linearly transformed by an operator $\mathcal{N}_{\otimes} : \boldsymbol{\mathcal{F}}_L \to \boldsymbol{\mathcal{F}}_L$ defined as

$$\mathcal{N}_{\otimes}(f) = \{(C^{\ell_1,\ell_2,\ell})^{\top}(\hat{f}_{t_1}^{\ell_1} \otimes \hat{f}_{t_2}^{\ell_2}) : \ell = 0, ..., L-1; (\ell_1, \ell_2) \in \mathbb{P}_L^{\ell}; t_1 = 0, ..., \tau_f^{\ell_1}; t_2 = 0, ..., \tau_f^{\ell_2}\}, \tag{10}$$

where for each degree $\ell \in \{0, ..., L-1\}$ the set

$$\mathbb{P}_L^{\ell} = \{(\ell_1, \ell_2) \in \{0, ..., L-1\}^2 : |\ell_1 - \ell_2| \leqslant \ell \leqslant \ell_1 + \ell_2\} \tag{11}$$

is defined in order to avoid the computation of trivially equivariant all-zero fragments. We make the dependence on $\mathbb{P}_L^{\ell}$ explicit since we redefine it in Section 3.

Unlike the pointwise activations discussed in the previous section this operator is strictly equivariant, with a mean relative equivariance error of 5.0×10^{-7} (see Appendix D). Note that applying this operator to signals on the sphere or rotation group results in generalized signals that are no longer on the sphere or rotation group. This is the rationale for the generalization to $\boldsymbol{\mathcal{F}}_L$: to unlock the ability to introduce non-linearity in a strictly equivariant manner. Note, however, that $g = \mathcal{N}_{\otimes}(f)$ has type $\tau_g = (\tau_g^0, ..., \tau_g^{L-1})$ where $\tau_g^{\ell} = \sum_{(\ell_1,\ell_2)\in\mathbb{P}_L^{\ell}} \tau_f^{\ell_1}\tau_f^{\ell_2}$ and therefore application of this non-linear operator results in a drastic expansion in representation size, which is problematic.

2.6 Generalized Spherical CNNs

Equipped with operators to both linearly and non-linearly transform elements of $\boldsymbol{\mathcal{F}}_L$, with the latter also performing degree mixing, we may consider a network with representation spaces $\boldsymbol{\mathcal{H}}^{(0)} = ... = \boldsymbol{\mathcal{H}}^{(S)} = \boldsymbol{\mathcal{F}}_L$. We consider the s-th layer of the network to take the form of a triple $\mathcal{A}^{(s)} = (\mathcal{L}_1, \mathcal{N}, \mathcal{L}_2)$ such that $\mathcal{A}^{(s)}(f^{(s-1)}) = \mathcal{L}_2(\mathcal{N}(\mathcal{L}_1(f^{(s-1)})))$, where $\mathcal{L}_1, \mathcal{L}_2 : \boldsymbol{\mathcal{F}}_L \to \boldsymbol{\mathcal{F}}_L$ are linear operators and $\mathcal{N} : \boldsymbol{\mathcal{F}}_L \to \boldsymbol{\mathcal{F}}_L$ is a non-linear activation operator.

The approaches of Cohen et al. (2018) and Esteves et al. (2018) are encompassed in this framework as $\mathcal{A}^{(s)} = (\mathcal{L}^{(\psi)}, \mathcal{N}_{\sigma}, \mathcal{I})$, where $\mathcal{I}$ denotes the identity operator and filters ψ may be defined to encode real-space properties such as localization (see Appendix C). The framework of Kondor et al. (2018) is also encompassed as $\mathcal{A}^{(s)} = (\mathcal{I}, \mathcal{N}_{\otimes}, \mathcal{L}_{\mathrm{G}}^{(\psi)})$. Here the generalized convolution $\mathcal{L}_{\mathrm{G}}^{(\psi)}$ comes last to counteract the representation-expanding effect of the tensor-product activation and prevent it from compounding as signals pass through the network. Appendix E lends intuition regarding relationships that may be captured by tensor-product activations followed by generalized convolutions.

For any intermediary representation $f^{(i)} \in \boldsymbol{\mathcal{F}}_L$ we may transition from equivariance with respect to the network input to invariance by discarding all but the scalar-valued fragments corresponding to $\ell = 0$ (equivalent to average pooling for signals on the sphere and rotation group). Finally, note that within this general framework we are free to consider hybrid approaches whereby layers proposed by Cohen et al. (2018); Esteves et al. (2018); Kondor et al. (2018) and others, and those presented subsequently, can be leveraged alongside each other within a single model.

3 Efficient Generalized Spherical CNNs

Existing approaches to spherical convolutional layers that are encompassed within the above framework are computationally demanding. They require the evaluation of costly harmonic transforms

on the sphere and rotation group. Furthermore, the only strictly rotationally equivariant non-linear layer is that of Kondor et al. (2018), which has an even greater computational cost, scaling with the fifth power of bandlimit — thereby limiting spatial resolution — and quadratically with the number of fragments per degree — thereby limiting representational capacity. This often prohibits the use of strictly equivariant spherical networks.

In this section we introduce a channel-wise structure, constrained generalized convolutions, and an optimized degree mixing set in order to construct new strictly equivariant layers that exhibit much improved scaling properties and parameter efficiency. Furthermore, we adopt efficient sampling theory on the sphere and rotation group to achieve additional computational savings.

3.1 Efficient Generalized Spherical Layers

For an activation $f \in \boldsymbol{\mathcal{F}}_L$ the value $\bar{\tau}_f = \frac{1}{L}\sum_{\ell=1}^{L-1} \tau_f^\ell$ represents a resolution-independent proxy for its representational capacity. Kondor et al. (2018) consider the separate fragments contained within f to subsume the traditional role of separate channels and therefore control the capacity of intermediary network representations through specification of τ_f. This is problematic because, whereas activation functions usually act on each channel separately and therefore have a cost that scales linearly with representational capacity (usually controlled by the number of channels), for the activation function $\mathcal{N}_\otimes$ not only does the cost scale quadratically with representational capacity $\bar{\tau}_f$, but so too does the size of $\mathcal{N}_\otimes(f)$. This feeds forward the quadratic dependence to the cost of, and number of parameters required by, the proceeding generalized convolution.

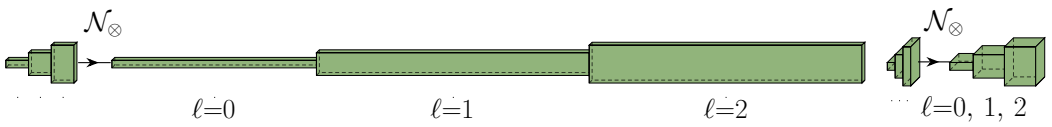
More specifically, note that computation of $g = \mathcal{N}_\otimes(f)$ requires the computation of $\sum_{\ell=0}^{L-1} \tau_g^\ell$ fragments, where $\tau_g^\ell = \sum_{(\ell_1,\ell_2)\in\mathbb{P}_L^\ell} \tau_f^{\ell_1}\tau_f^{\ell_2}$. The size of $\mathbb{P}_L^\ell$ is $\mathcal{O}(L\ell)$ for each ℓ and therefore the expanded representation has size $\sum_{\ell=0}^{L-1} \tau_g^\ell$, of order $\mathcal{O}(\bar{\tau}_f^2 L^3)$. By exploiting the sparsity of Clebsch-Gordan coefficients ($C_{m_1,m_2,m}^{\ell_1,\ell_2,\ell} = 0$ if $m_1 + m_2 \neq m$) each fragment $(C^{\ell_1,\ell_2,\ell})^\top(\hat{f}^{\ell_1} \otimes \hat{f}^{\ell_2})$ can be computed in $\mathcal{O}(\ell \min(\ell_1, \ell_2))$. Hence, the total cost of computing all necessary fragments has complexity $\mathcal{O}(C^2L^5)$, where $C = \bar{\tau}_f$ captures representational capacity.

3.1.1 Channel-Wise Tensor-Product Activations

As is more standard for CNNs we maintain a separate channels axis, with network activations taking the form $(f_1, ..., f_K) \in \boldsymbol{\mathcal{F}}_L^K$ where $f_i \in \boldsymbol{\mathcal{F}}_L$ all share the same type τ_f. The non-linearity $\mathcal{N}_\otimes$ may then be applied to each channel separately at a cost that is reduced by K-times relative to its application on a single channel with the same total number of fragments. This saving arises since for each ℓ we need only compute $K\sum_{(\ell_1,\ell_2)\in\mathbb{P}_L^\ell} \tau_f^{\ell_1}\tau_f^{\ell_2}$ fragments rather than $\sum_{(\ell_1,\ell_2)\in\mathbb{P}_L^\ell} (K\tau_f^{\ell_1})(K\tau_f^{\ell_2})$.

Figure 1 visualizes this reduction for the case $K = 3$. Note, however, that for practical applications $K \sim 100$ is more typical. The K-times reduction in cost is therefore substantial and allows for intermediary activations with orders of magnitude more representational capacity.

By introducing this multi-channel approach and using $C = K$ rather than $C = \bar{\tau}_f$ to control representational capacity, we reduce the complexity of $\mathcal{N}_\otimes$ with respect to representational capacity from $\mathcal{O}(C^2)$ to $\mathcal{O}(C)$.



(a) Prior approach to applying a tensor-product based non-linear operator (b) Ours

Figure 1: Comparison (to scale) of the expansion caused by the tensor-product activation applied to inputs of equal representational capacity but different structure. With depth representing the number of channels and width the number of fragments for each degree, it is clear that by grouping fragments into K separate channels the expansion (and therefore cost) can be K-times reduced. Visualization corresponds to inputs with (L, K) equal to $(3, 1)$ and $(3, 3)$ for panel (a) and (b), respectively.

3.1.2 CONSTRAINED GENERALIZED CONVOLUTION

Although much reduced, for a signal $f \in \boldsymbol{\mathcal{F}}_L^{K_{\text{in}}}$ the channel-wise application of $\mathcal{N}_\otimes$ still results in a drastically expanded representation $g = \mathcal{N}_\otimes(f)$, to which a representation-contracting generalized convolution must be applied in order to project onto a new activation $g' = \mathcal{L}_G^{(\psi)}(g) \in \boldsymbol{\mathcal{F}}_L^{K_{\text{out}}}$ of the desired type $\tau_{g'}$ and number of channels K_{out}. However, under our multi-channel structure computational and parameter efficiency can be improved significantly by decomposing $\mathcal{L}_{\text{G}}^{(\psi)}$ into three separate linear operators, $\mathcal{L}_{\text{G}_1}^{(\psi_1)}$, $\mathcal{L}_{\text{G}_2}^{(\psi_2)}$ and $\mathcal{L}_{\text{G}_3}^{(\psi_3)}$.

The first, $\mathcal{L}_{\text{G}_1}^{(\psi_1)}$, acts uniformly across channels, performing a linear projection down onto the desired type, and should be interpreted as a learned extension of $\mathcal{N}_\otimes$ which undoes the drastic expansion. The second, $\mathcal{L}_{\text{G}_2}^{(\psi_2)}$, then acts channel-wise, taking linear combinations of the (contracted number of) fragments within each channel. The third, $\mathcal{L}_{\text{G}_3}^{(\psi_3)}$, acts across channels, taking linear combinations to learn new features. More concretely, the three filters are of the form $\psi_1 = \{\hat{\psi}_1^\ell \in \mathbb{C}^{\tau_g^\ell \times \tau_{g'}^\ell} : \ell = 0, ..., L-1\}$, $\psi_2 = \{\hat{\psi}_2^{\ell,k} \in \mathbb{C}^{\tau_{g'}^{\ell} \times \tau_{g'}^\ell} : \ell = 0, ..., L-1; k = 1, ..., K_{\text{in}}\}$ and $\psi_3 = \{\hat{\psi}_3^\ell \in \mathbb{C}^{K_{\text{in}} \times K_{\text{out}}} : \ell = 0, ..., L-1\}$, rather than a single filter of the form $\psi = \{\hat{\psi}^\ell \in \mathbb{C}^{K_{\text{in}} \times \tau_g^\ell \times K_{\text{out}} \times \tau_{g'}^\ell} : \ell = 0, ..., L-1\}$, leading to a large reduction in the number of parameters as τ_g^ℓ is invariably very large.

By applying the first step uniformly across channels we minimize the parametric dependence on the expanded representation and allow new features to be subsequently learned much more efficiently. Together the second and third steps can be seen as analogous to depthwise separable convolutions often used in planar convolutional networks.

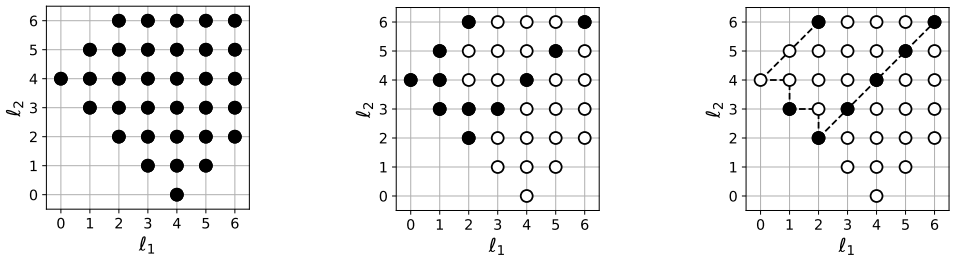
3.1.3 OPTIMIZED DEGREE MIXING SETS

We now consider approaches to reduce the $\mathcal{O}(L^5)$ complexity with respect to spatial resolution L. In the definition of $\mathcal{N}_\otimes$ each element of $\mathbb{P}_L^\ell$ independently defines an equivariant fragment. Therefore a restricted $\mathcal{N}_\otimes$ in which only a subset of $\mathbb{P}_L^\ell$ is used for each degree ℓ still defines a strictly equivariant operator, while reducing computational complexity. In order to make savings whilst remaining at resolution L it is necessary to consider subsets of $\mathbb{P}_L^\ell$ that scale better than $\mathcal{O}(L^2)$. The challenge is to find such subsets that do not hamper the ability of the resulting operator to inject non-linearity and mix information corresponding to different degrees ℓ.

Whilst various subsetting approaches are possible, the following argument motivates an approach that we have found to be particularly effective. If $(\ell_1, \ell_3) \in \mathbb{P}_L^\ell$, then representational space is designated to capture the relationship between ℓ_1 and ℓ_3-degree information. However, if resources have been designated already to capture the relationship between ℓ_1 and ℓ_2-degrees, as well as between ℓ_2 and ℓ_3-degrees, then some notion of the relationship between ℓ_1 and ℓ_3-degrees has been captured already. Consequently, it is unnecessary to designate further resources for this purpose.

More generally, consider the graph $G_L^\ell = (\mathbb{N}_L, \mathbb{P}_L^\ell)$ with nodes $\mathbb{N}_L = \{0, ..., L-1\}$ and edges $\mathbb{P}_L^\ell$. A restricted tensor-product activation can be constructed by using a subset of $\mathbb{P}_L^\ell$ that corresponds to a subgraph of G_L^ℓ. The subgraph of G_L^ℓ captures some notion of the relationship between incoming ℓ_1 and ℓ_2-degree information if it contains a path between nodes ℓ_1 and ℓ_2. Therefore we are interested in subgraphs for which there exists a path between any two nodes if there exists such a path in the original graph, guaranteeing that any degree-mixing relationship captured by the original graph is also captured by the subgraph.

The smallest subgraph satisfying this property is a minimum spanning tree (MST) of G_L^ℓ. The set of edges corresponding to any MST has at most L elements and we choose to consider its union with the set of loop-edges in G_L^ℓ (of the form (ℓ_1, ℓ_1)), which proved particularly important for injecting non-linearity. We denote the resulting set as $\bar{\mathbb{P}}_L^\ell$ and note that it satisfies $|\bar{\mathbb{P}}_L^\ell| \leqslant 2L$. Therefore the tensor-product activation $\bar{\mathcal{N}}_\otimes$ corresponding to Equation 11 with $\mathbb{P}_L^\ell$ replaced by $\bar{\mathbb{P}}_L^\ell$ has reduced spatial complexity $\mathcal{O}(L^4)$. Given that many minimal spanning trees of the unweighted graph G_L^ℓ exist for each ℓ, we select the ones that minimize the cost of the resulting activation $\bar{\mathcal{N}}_\otimes$ by assigning to each edge (ℓ_1, ℓ_2) in G_L^ℓ a weight equal to the cost of computing $(C^{\ell_1,\ell_2,\ell})^\top(\hat{f}^{\ell_1} \otimes \hat{f}^{\ell_2})$ and selecting the MST of the weighted graph.



(a) Full $\mathbb{P}_L^\ell$ set of size $\mathcal{O}(L^2)$ (b) MST subset of size $\mathcal{O}(L)$ (c) RMST subset of size $\mathcal{O}(\log L)$

Figure 2: Visualization of the mixing set $\mathbb{P}_L^\ell$ (for $L = 7$ and $\ell = 4$) and the approaches to subsetting based on the minimum spanning tree (MST) and reduced minimum spanning tree (RMST) mixing polices, which reduces related computation costs from $\mathcal{O}(L^2)$ to, respectively, $\mathcal{O}(L)$ or $\mathcal{O}(\log L)$.

An example of $\mathbb{P}_L^\ell$ and a MST subset $\bar{\mathbb{P}}_L^\ell$ is shown in Figure 2, where the dashed line in Figure 2c shows the general form of the MST. Using this as a principled starting point we consider the further reduced MST (RMST) subset $\tilde{\mathbb{P}}_L^\ell$ corresponding to centering the MST at the edge (ℓ, ℓ) and retaining only the edges that fall a distance of 2^i away on the dotted line for some $i \in \mathbb{N}$. We use $\tilde{\mathcal{N}}_\otimes$ to denote the corresponding operator and note that it has further reduced spatial complexity of $\mathcal{O}(L^3 \log L)$.

We demonstrate in Section 4 that networks that make use of the MST tensor-product activation achieve state-of-the-art performance. Replacing the MST with RMST activation results in a small but insignificant degradation in performance, which is offset by the reduced computational cost.

3.1.4 Reduction in Computational and Memory Footprints

The three modifications proposed in Sections 3.1.1 to 3.1.3 are motivated by improved scaling properties. Importantly, they also translate to large reductions in the computational and memory cost of strictly equivariant layers in practice, as detailed in Appendix F. Even at a modest bandlimit of $L = 64$ and relatively small number of channels $K = 4$, for example, the modifications together lead to a 51-times reduction in the number of flops required for computations and 16-times reduction in the amount of memory required to store representations, weights and gradients for training.

3.2 Efficient Sampling Theory

By adopting sampling theorems on the sphere we provide access to underlying continuous signal representations that fully capture the symmetries and geometric properties of the sphere, and allow standard convolutions to be computed exactly and efficiently through their harmonic representations, as discussed in greater detail in Appendices A and B. We adopt the efficient sampling theorems on sphere and rotation group of McEwen & Wiaux (2011) and McEwen et al. (2015a), respectively, which reduce the Nyquist rate by a factor of two compared to those of Driscoll & Healy (1994) and Kostelec & Rockmore (2008), which have been adopted in other spherical CNN constructions (e.g. Cohen et al., 2018; Kondor et al., 2018; Esteves et al., 2018; 2020). The sampling theorems adopted are equipped with fast algorithms to compute harmonic transforms, with complexity $\mathcal{O}(L^3)$ for the sphere and $\mathcal{O}(L^4)$ for the rotation group. When imposing an azimuthal bandlimit $N \ll L$, the complexity of transforms on the rotation group can be reduced to $\mathcal{O}(NL^3)$, which we often exploit in our standard (non-generalized) convolutional layers.

4 Experiments

Using our efficient generalized spherical CNN framework (implemented in the `fourpiAI` [1] code) we construct networks that we apply to numerous spherical benchmark problems. We achieve state-of-the-art performance, demonstrating enhanced equivariance without compromising representational capacity or parameter efficiency. In all experiments we use a similar architecture, consisting of 2–3 standard convolutional layers (e.g. $\mathbb{S}^2$ or $\mathrm{SO}(3)$ convolutions proceeded by ReLUs), followed by 2–3 of our efficient generalized layers. We adopt the efficient sampling theory described in Section 3.2 and encode localization of spatial filters as discussed in Appendix C. Full experimental details may be found in Appendix G.

[1] `https://www.kagenova.com/products/fourpiAI/`

Table 1: Test accuracy for spherical MNIST digits classification problem

	NR/NR	R/R	NR/R	Params
Planar CNN	99.32	90.74	11.36	58k
Cohen et al. (2018)	95.59	94.62	93.40	58k
Kondor et al. (2018)	96.40	96.60	96.00	286k
Esteves et al. (2020)	**99.37**	99.37	99.08	58k
Ours (MST)	99.35	**99.38**	**99.34**	58k
Ours (RMST)	99.29	99.17	99.18	**57k**

Table 2: Test root mean squared (RMS) error for QM7 regression problem

	RMS	Params
Montavon et al. (2012)	5.96	-
Cohen et al. (2018)	8.47	1.4M
Kondor et al. (2018)	7.97	>1.1M
Ours (MST)	**3.16**	337k
Ours (RMST)	3.46	**335k**

Table 3: SHREC'17 object retrieval competition metrics (perturbed micro-all)

	P@N	R@N	F1@N	mAP	NDCG	Params
Kondor et al. (2018)	0.707	0.722	0.701	0.683	0.756	>1M
Cohen et al. (2018)	0.701	0.711	0.699	0.676	0.756	1.4M
Esteves et al. (2018)	0.717	**0.737**	-	**0.685**	-	500k
Ours	**0.719**	0.710	**0.708**	0.679	**0.758**	**250k**

4.1 Rotated MNIST on the Sphere

We consider the now standard benchmark problem of classifying MNIST digits projected onto the sphere. Three experimental modes NR/NR, R/R and NR/R are considered, indicating whether the training/test sets have been randomly rotated (R) or not (NR). Results are presented in Table 1, which shows that we closely match the prior state-of-the-art performance obtained by Esteves et al. (2020) on the NR/NR and R/R modes, whilst outperforming all previous spherical CNNs on the NR/R mode, demonstrating the increased degree of equivariance achieved by our model.

Results are shown for models using both the MST-based and RMST-based mixing sets within the tensor-product activation. The results obtained when using the full sets $\mathbb{P}_L^\ell$ are very similar to those obtained when using the MST-based sets (e.g. full sets achieved an accuracy of 99.39 for R/R).

4.2 Atomization Energy Prediction

We consider the problem of regressing the atomization energy of molecules given the molecule's Coulomb matrix and the positions of the atoms in space, using the QM7 dataset (Blum & Reymond, 2009; Rupp et al., 2012). Results are presented in Table 2, which shows that we dramatically outperform other approaches, whilst using significantly fewer parameters.

4.3 3D Shape Retrieval

We consider the 3D shape retrieval problem on the SHREC'17 (Savva et al., 2017) competition dataset, containing 51k 3D object meshes. We follow the pre-processing step of Cohen et al. (2018), where several spherical projections of each mesh are computed, and use the official SHREC'17 data splits. Results are presented in Table 3 for the standard SHREC precision and recall metrics, which shows that we achieve state-of-the-art performance compared to other spherical CNN approaches, achieving the highest three of five performance metrics, whilst using significantly fewer parameters.

5 Conclusions

We have presented a generalized framework for CNNs on the sphere that encompasses various existing approaches. We developed new efficient layers to be used as primary building blocks in this framework by introducing a channel-wise structure, constrained generalized convolutions, and optimized degree mixing sets determined by minimum spanning trees. These new efficient layers exhibit strict rotational equivariance, without compromising on representational capacity or parameter efficiency. When combined with the flexibility of the generalized framework to leverage the strengths of alternative layers, powerful hybrid models can be constructed. On all spherical benchmark problems considered we achieve state-of-the-art performance, both in terms of accuracy and parameter efficiency. In future work we intend to improve the scalability of our generalized framework further still. In particular, we plan to introduce additional highly scalable layers, for example by extending scattering transforms (Mallat, 2012) to the sphere, to further realize the potential of deep learning on a host of new applications where spherical data are prevalent.

REFERENCES

Lorenz Blum and Jean-Louis Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database GDB-13. *Journal of the American Chemical Society*, 131:8732, 2009.

Wouter Boomsma and Jes Frellsen. Spherical convolutions and their application in molecular modelling. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 3433–3443. Curran Associates, Inc., 2017.

Taco Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1801.10130.

Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. *arXiv preprint arXiv:1902.04615*, 2019. URL https://arxiv.org/abs/1902.04615.

James Driscoll and Dennis Healy. Computing Fourier transforms and convolutions on the sphere. *Advances in Applied Mathematics*, 15:202–250, 1994.

Carlos Esteves. Theoretical aspects of group equivariant neural networks. *arXiv preprint arXiv:2004.05154*, 2020. URL https://arxiv.org/abs/2004.05154.

Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning SO(3) equivariant representations with spherical CNNs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 52–68, 2018. URL https://arxiv.org/abs/1711.06721.

Carlos Esteves, Ameesh Makadia, and Kostas Daniilidis. Spin-weighted spherical CNNs. *arXiv preprint arXiv:2006.10731*, 2020. URL https://arxiv.org/abs/2006.10731.

Jean Gallier and Jocelyn Quaintance. *Aspects of Harmonic Analysis and Representation Theory*. 2019. URL https://www.seas.upenn.edu/~jean/nc-harmonic.pdf.

Dennis Healy, Daniel Rockmore, Peter Kostelec, and S. Moore. FFTs for the 2-sphere – improvements and variations. *Journal of Fourier Analysis and Applications*, 9(4):341–385, 2003.

Chiyu Jiang, Jingwei Huang, Karthik Kashinath, Philip Marcus, Matthias Niessner, et al. Spherical CNNs on unstructured grids. *arXiv preprint arXiv:1901.02039*, 2019. URL https://arxiv.org/abs/1901.02039.

Rodney A Kennedy and Parastoo Sadeghi. *Hilbert space methods in signal processing*. Cambridge University Press, 2013.

Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic gradient descent. In *ICLR: International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, pp. 2747–2755, 2018. URL https://arxiv.org/abs/1802.03690.

Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch-Gordan nets: a fully fourier space spherical convolutional neural network. In *Advances in Neural Information Processing Systems*, pp. 10117–10126, 2018. URL https://arxiv.org/abs/1806.09231.

Peter Kostelec and Daniel Rockmore. FFTs on the rotation group. *Journal of Fourier Analysis and Applications*, 14:145–179, 2008.

Stéphane Mallat. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012. URL https://arxiv.org/abs/1101.2286.

Domenico Marinucci and Giovanni Peccati. *Random Fields on the Sphere: Representation, Limit Theorem and Cosmological Applications*. Cambridge University Press, 2011.

Jason McEwen and Yves Wiaux. A novel sampling theorem on the sphere. *IEEE Transactions on Signal Processing*, 59(12):5876–5887, 2011. URL https://arxiv.org/abs/1110.6298.

Jason McEwen, Michael P. Hobson, Daniel J. Mortlock, and Anthony N. Lasenby. Fast directional continuous spherical wavelet transform algorithms. *IEEE Trans. Sig. Proc.*, 55(2):520–529, 2007. URL https://arxiv.org/abs/astro-ph/0506308.

Jason McEwen, Pierre Vandergheynst, and Yves Wiaux. On the computation of directional scale-discretized wavelet transforms on the sphere. In *Wavelets and Sparsity XV, SPIE international symposium on optics and photonics, invited contribution*, volume 8858, 2013. URL https://arxiv.org/abs/1308.5706.

Jason McEwen, Martin Büttner, Boris Leistedt, Hiranya V Peiris, and Yves Wiaux. A novel sampling theorem on the rotation group. *IEEE Signal Processing Letters*, 22(12):2425–2429, 2015a. URL https://arxiv.org/abs/1508.03101.

Jason McEwen, Boris Leistedt, Martin Büttner, Hiranya Peiris, and Yves Wiaux. Directional spin wavelets on the sphere. *IEEE Trans. Sig. Proc., submitted*, 2015b. URL https://arxiv.org/abs/1509.06749.

Jason McEwen, Claudio Durastanti, and Yves Wiaux. Localisation of directional scale-discretised wavelets on the sphere. *Applied Comput. Harm. Anal.*, 44(1):59–88, 2018. URL https://arxiv.org/abs/1509.06767.

Grégoire Montavon, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole V. Lilienfeld, and Klaus-Robert Müller. Learning invariant representations of molecules for atomization energy prediction. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25*, pp. 440–448. Curran Associates, Inc., 2012.

Nathanaël Perraudin, Michaël Defferrard, Tomasz Kacprzak, and Raphael Sgier. Deepsphere: Efficient spherical convolutional neural network with HEALPix sampling for cosmological applications. *Astronomy and Computing*, 27:130–146, 2019. URL https://arxiv.org/abs/1810.12186.

Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108:058301, 2012. URL https://arxiv.org/abs/1109.2618.

Manolis Savva, Fisher Yu, Hao Su, Asako Kanezaki, Takahiko Furuya, Ryutarou Ohbuchi, Zhichao Zhou, Rui Yu, Song Bai, Xiang Bai, et al. Large-scale 3d shape retrieval from shapenet core55: Shrec'17 track. In *Proceedings of the Workshop on 3D Object Retrieval*, pp. 39–50. Eurographics Association, 2017.

Max Tegmark. An Icosahedron-Based method for pixelizing the celestial sphere. *Astrophys. J. Lett.*, 470:L81, October 1996. URL https://arxiv.org/abs/astro-ph/9610094.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. URL https://arxiv.org/abs/1802.08219.

Stefano Trapani and Jorge Navaza. Calculation of spherical harmonics and Wigner *d* functions by FFT. Applications to fast rotational matching in molecular replacement and implementation into *AMoRe*. *Acta Crystallographica Section A*, 62(4):262–269, 2006.

Benjamin Wandelt and Krzysztof Górski. Fast convolution on the sphere. *Phys. Rev. D.*, 63(12): 123002, 2001. URL https://arxiv.org/abs/astro-ph/0008227.

A REPRESENTATIONS OF SIGNALS ON THE SPHERE AND ROTATION GROUP

To provide further context for the discussion presented in the introduction and to elucidate the properties of different sampling theory on the sphere and rotation group, we concisely review representations of signals on the sphere and rotation group.

A.1 DISCRETIZATION

It is well-known that a completely regular point distribution on the sphere does in general not exist (e.g. Tegmark, 1996). Consequently, while a variety of spherical discretization schemes exists (e.g. icosahedron, HEALPix, graph, and other representations), it is not possible to discretize (i.e. to sample or pixelize) the sphere in a manner that is invariant to rotations, i.e. a discrete sampling of rotations of the samples on the sphere will in general not map onto the same set of sample positions. This differs to the Euclidean setting and has important implications when constructing convolution operators on the sphere, which clearly are a critical component of CNNs.

Since convolution operators are in general built using a translation operator – equivalently a rotation operator when on the sphere – it is thus not possible to construct a convolution operator directly on a discretized representation of the sphere that captures all of the symmetries of the underlying spherical manifold. While approximate discrete representations can be considered, and are nevertheless useful, such representations cannot capture all underlying spherical symmetries.

A.2 SAMPLING THEORY

Alternative representations, however, can capture all underlying spherical symmetries. Sampling theories on the sphere (e.g. Driscoll & Healy, 1994; McEwen & Wiaux, 2011) provide a mechanism to capture all information content of an underlying continuous function on the sphere from a finite set of samples (and similarly on the rotation group; Kostelec & Rockmore 2008; McEwen et al. 2015a). A sampling theory on the sphere is equivalent to a cubature (i.e. quadrature) rule for the exact integration of a bandlimited functions on the sphere. While optimal cubature on the sphere remains an open problem, the most efficient sampling theory on the sphere and rotation group is that developed by McEwen & Wiaux (2011) and McEwen et al. (2015a), respectively.

On a compact manifold like the sphere (and rotation group), harmonic (i.e. Fourier) space is discrete. Hence, a finite set of harmonic coefficients captures all information content of an underlying continuous bandlimited signal. Since such a representation provides access to the underlying continuous signal, all symmetries and geometric properties of the sphere are captured perfectly. Such representations have been employed extensively in the construction of wavelet transforms on the sphere, where the use of sampling theorems on the sphere and rotation group yield wavelet transforms of discretized continuous signals that are theoretically exact (e.g. McEwen et al., 2013; 2015b; 2018). Harmonic signal representations have also been exploited in spherical CNNs to access all underlying spherical symmetries and develop equivariance network layers (Cohen et al., 2018; Kondor et al., 2018; Esteves et al., 2018; 2020).

A.3 EXACT AND EFFICIENT COMPUTATION

Signals on the sphere $f \in L^2(\mathbb{S}^2)$ may be decomposed into their harmonic representations as

$$f(\omega) = \sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{\ell} f_m^\ell Y_m^\ell(\omega), \tag{12}$$

where their spherical harmonic coefficients are given by

$$f_m^\ell = \langle f, Y_m^\ell \rangle = \int_{\mathbb{S}^2} \mathrm{d}\mu(\omega) f(\omega)\, Y_m^{\ell *}(\omega), \tag{13}$$

for $\omega \in \mathbb{S}^2$. Similarly, signals on the rotation group $g \in L^2(\mathrm{SO}(3))$ may be decomposed into their harmonic representations as

$$g(\rho) = \sum_{\ell=0}^{\infty} \frac{2\ell+1}{8\pi^2} \sum_{m=-\ell}^{\ell} \sum_{n=-\ell}^{\ell} g_{mn}^\ell D_{mn}^{\ell *}(\rho) \tag{14}$$

where their harmonic (Wigner) coefficients are given by

$$g^\ell_{mn} = \langle g, D^{\ell *}_{mn} \rangle = \int_{\mathrm{SO}(3)} \mathrm{d}\mu(\rho) g(\rho)\, D^\ell_{mn}(\rho), \tag{15}$$

for $\rho \in \mathrm{SO}(3)$. Note that we adopt the convention where the conjugate of the Wigner D-function is used in Equation 14 since this leads to a convenient harmonic representation when considering convolutions (cf. McEwen et al., 2015a; 2018).

As mentioned above, sampling theory pertains to strategies to capture all of the information content of band limited signals from a finite set of samples. Since the harmonic space of the sphere and rotation group is discrete, this is equivalent to an exact quadrature rule for the computation of harmonic coefficients by Equation 13 and Equation 15 from sampled signals.

The canonical equiangular sampling theory on the sphere was that developed by Driscoll & Healy (1994), and subsequently extended to the rotation group by Kostelec & Rockmore (2008). More recently, novel sampling theorems on the sphere and rotation group were developed by McEwen & Wiaux (2011) and McEwen et al. (2015a), respectively, that reduce the Nyquist rate by a factor of two. Previous CNN constructions on the sphere (e.g. Cohen et al., 2018; Kondor et al., 2018; Esteves et al., 2018; 2020) have adopted the more well-known sampling theories of Driscoll & Healy (1994) and Kostelec & Rockmore (2008). In contrast, we adopt the more efficient sampling theories of McEwen & Wiaux (2011) and McEwen et al. (2015a) to provide additional efficiency savings, implemented in the open source `ssht` [2] and `so3` [3] software packages (we also make use of a TensorFlow implementation of these algorithms in our private `tensossht` [4] code). Note also that the sampling schemes associated with the theory of McEwen & Wiaux (2011) (and other minor variants implemented in `ssht`) align more closely with the one-to-two aspect ratio of common spherical data, such as 360° photos and videos.

All of the sampling theories discussed are equipped with fast algorithms to compute harmonic transforms, with complexity $\mathcal{O}(L^3)$ for transforms on the sphere (Driscoll & Healy, 1994; McEwen & Wiaux, 2011) and complexity $\mathcal{O}(L^4)$ for transforms on the rotation group (Kostelec & Rockmore, 2008; McEwen et al., 2015a). Note that algorithms that achieve slightly lower complexity have been developed (Driscoll & Healy, 1994; Healy et al., 2003; Kostelec & Rockmore, 2008) but these are known to suffer stability issues (Healy et al., 2003; Kostelec & Rockmore, 2008). By imposing an azimuthally bandlimit N, where typically $N \ll L$, the complexity of transforms on the rotation group can be reduced to $\mathcal{O}(NL^3)$ (McEwen et al., 2015a), which we exploit in our networks.

These fast algorithms to compute harmonic transforms on the sphere and rotation group can be leveraged to yield the exact and efficient computation of convolutions through their harmonic representations (see Appendix B). By computing convolutions in harmonic space, pixelization and quadrature errors are avoided and computational complexity is reduced to the cost of the respective harmonic transforms.

B Convolution on the Sphere and Rotation Group

For completeness we make explicit the standard (non-generalized) convolution operations on the sphere and rotation group that we adopt. The general form of convolution for signals $f \in L^2(\Omega)$ either on the sphere ($\Omega = \mathbb{S}^2$) or rotation group ($\Omega = \mathrm{SO}(3)$) is specified by Equation 1, with harmonic representation given by Equation 3. Here we provide specific expressions for the convolution for a variety of cases, describe the normalization constants that arise and may be absorbed into learnable filters, and derive the corresponding harmonic forms. In practice all convolutions are computed in harmonic space since the computation is then exact, avoiding pixelisation or quadrature errors, and efficient when fast algorithms to compute harmonic transforms are exploited (see Appendix A).

[2] `http://www.spinsht.org/`

[3] `http://www.sothree.org/`

[4] Available on request from `https://www.kagenova.com/`.

B.1 CONVOLUTION ON THE SPHERE

Given two spherical signals $f, \psi \in L^2(\mathbb{S}^2)$ their convolution, which in general is a signal on the rotation group, may be decomposed as

$$
\begin{aligned}
(f \star \psi)(\rho) &= \langle f, R_\rho \psi \rangle && (16)\\
&= \int_{S^2} d\Omega(\omega) f(\omega) \psi^*(\rho^{-1}\omega) && (17)\\
&= \sum_{\ell m} \sum_{\ell' m'} \sum_{n} f_m^\ell D_{m'n}^{\ell'*}(\rho) \psi_n^{\ell'*} \int_{S^2} d\Omega(\omega) Y_m^\ell(\omega) Y_{m'}^{\ell'*}(\omega) && (18)\\
&= \sum_{\ell m} \sum_{\ell' m'} \sum_{n} f_m^\ell D_{m'n}^{\ell'*}(\rho) \psi_n^{\ell'*} \delta_{\ell\ell'} \delta_{mm'} && (19)\\
&= \sum_{\ell m n} \left(f_m^\ell \psi_n^{\ell *} \right) D_{mn}^{\ell *}(\rho), && (20)
\end{aligned}
$$

yielding harmonic coefficients

$$
(f * \psi)_{mn}^\ell = \frac{8\pi^2}{2\ell + 1} f_m^\ell \psi_n^{\ell *}. \tag{21}
$$

The constants $8\pi^2/(2\ell+1)$ may be absorbed into learnable parameters.

B.2 CONVOLUTION ON THE SPHERE WITH AXISYMMETRIC FILTERS

When convolving a spherical signal $f \in L^2(\mathbb{S}^2)$ with an axisymmetric spherical filter $\psi \in L^2(\mathbb{S}^2)$ that is invariant to azimuthal rotations, the resultant $(f \star \psi)$ may be interpreted as a signal on the sphere. To see this note that an axisymmetric filter ψ has harmonic coefficients $\psi_n^\ell = \psi_0^\ell \delta_{n0}$ that are non-zero only for $m = 0$. Denoting rotations by their zyz-Euler angles $\rho = (\alpha, \beta, \gamma)$ and substituting into Equation 20 we see that the convolution may be decomposed as

$$
\begin{aligned}
(f \star \psi)(\alpha, \beta, \gamma) &= \sum_{\ell m n} \left(f_m^\ell \psi_0^{\ell *} \delta_{n0} \right) D_{mn}^{\ell *}(\alpha, \beta, \gamma) && (22)\\
&= \sum_{\ell m} f_m^\ell \psi_0^{\ell *} D_{m0}^{\ell *}(\alpha, \beta, 0) && (23)\\
&= \sum_{\ell m} f_m^\ell \psi_0^{\ell *} \sqrt{\frac{4\pi}{2\ell + 1}} Y_m^\ell(\beta, \alpha). && (24)
\end{aligned}
$$

We may therefore interpret $(f \star \psi)$ as a signal on the sphere with spherical harmonic coefficients

$$
(f \star \psi)_m^\ell = \sqrt{\frac{4\pi}{2\ell + 1}} f_m^\ell \psi_0^{\ell *}. \tag{25}
$$

The constants $\sqrt{4\pi/(2\ell+1)}$ may be absorbed into learnable parameters.

B.3 CONVOLUTION ON THE ROTATION GROUP

Given two signals $f, \psi \in L^2(\mathrm{SO}(3))$ on the rotation group their convolution may then be decomposed as

$$(f \star \psi)(\rho) = \langle f, R_\rho \psi \rangle \tag{26}$$

$$= \int_{\mathrm{SO}(3)} d\mu(\rho') f(\rho') \psi^*(\rho^{-1}\rho') \tag{27}$$

$$= \int_{\mathrm{SO}(3)} d\mu(\rho') \Big[\sum_\ell \frac{2\ell+1}{8\pi^2} \sum_{mn} f^\ell_{mn} D^{\ell *}_{mn}(\rho') \Big] \Big[\sum_{\ell'} \frac{2\ell'+1}{8\pi^2} \sum_{m'n'} \psi^{\ell' *}_{m'n'} D^{\ell'}_{m'n'}(\rho^{-1}\rho) \Big] \tag{28}$$

$$= \sum_\ell \frac{2\ell+1}{8\pi^2} \sum_{mn} f^\ell_{mn} \sum_{\ell'} \frac{2\ell'+1}{8\pi^2} \sum_{m'n'} \psi^{\ell' *}_{m'n'} \int_{\mathrm{SO}(3)} d\mu(\rho') D^{\ell *}_{mn}(\rho') D^{\ell'}_{m'n'}(\rho^{-1}\rho') \tag{29}$$

$$= \sum_\ell \frac{2\ell+1}{8\pi^2} \sum_{mn} f^\ell_{mn} \sum_{\ell'} \frac{2\ell'+1}{8\pi^2} \sum_{m'n'} \psi^{\ell' *}_{m'n'} \int_{\mathrm{SO}(3)} d\mu(\rho') D^{\ell *}_{mn}(\rho') \sum_k D^{\ell' *}_{km'}(\rho) D^{\ell'}_{kn'}(\rho') \tag{30}$$

$$= \sum_\ell \frac{2\ell+1}{8\pi^2} \sum_{mn} f^\ell_{mn} \sum_{\ell'} \frac{2\ell'+1}{8\pi^2} \sum_{m'n'} \psi^{\ell' *}_{m'n'} \sum_k D^{\ell' *}_{km'}(\rho) \frac{8\pi^2}{2\ell+1} \delta_{\ell\ell'} \delta_{mk} \delta_{nn'} \tag{31}$$

$$= \sum_{\ell m m'} \frac{2\ell+1}{8\pi^2} D^{\ell *}_{mm'}(\rho) \Big(\sum_n f^\ell_{mn} \psi^{\ell *}_{m'n} \Big), \tag{32}$$

where for Equation 30 we make use of the relation (e.g. Marinucci & Peccati, 2011; McEwen et al., 2018)

$$D^\ell_{mn}(\rho^{-1}\rho') = \sum_k D^{\ell *}_{km}(\rho) D^\ell_{kn}(\rho'). \tag{33}$$

This decomposition yields harmonic coefficients

$$(f * \psi)^\ell_{mn} = \sum_{m'} f^\ell_{mm'} \psi^{\ell *}_{nm'}. \tag{34}$$

C FILTERS ON THE SPHERE AND ROTATION GROUP

When defining filters we look to encode desirable real-space properties, such as locality and regularity. However, in practice considerable computation may be saved by defining the filters in harmonic space and saving the cost of harmonic transforming ahead of harmonic space convolutions. We describe here how filters motivated by their real space properties may be defined directly in harmonic space.

C.1 DIRAC DELTA FILTERS ON THE SPHERE

Spherical filters may be constructed as a weighted sum of Dirac delta functions on the sphere. This construction is useful as the harmonic representation has an analytic form that may be computed efficiently. Furthermore, various real space properties can be encoded through sensible placement of the Dirac delta functions.

The spherical Dirac delta function $\delta_{\omega'}$ centered at $\omega' = (\theta', \phi') \in \mathbb{S}^2$ is defined as

$$\delta_{\omega'}(\omega) = \frac{1}{\sin\theta} \delta_{\mathbb{R}}(\cos\theta - \cos\theta') \delta_{\mathbb{R}}(\phi - \phi'), \tag{35}$$

where $\delta_{\mathbb{R}}$ is the familiar Dirac delta function on the reals centered at 0. The Dirac delta on the sphere may be represented in harmonic space by

$$(\delta_{\omega'})^\ell_m = Y^{\ell *}_m(\omega') = N^\ell_m P^\ell_m(\cos\theta') e^{-im\phi'}, \tag{36}$$

which follows form the sifting property of the Dirac delta, and where Y_m^ℓ denote the spherical harmonic functions, $P_m^\ell(x)$ are associated Legendre functions and

$$N_m^\ell = \sqrt{\frac{2\ell+1}{4\pi}\frac{(l-m)!}{(l+m)!}} \tag{37}$$

is a normalizing constant.

This representation may then be used to define a filter $\psi \in L^2(\mathbb{S}^2)$ as a weighted sum of spherical Dirac delta functions, with weights w_{ij} assigned to Dirac delta functions centered at points $\{(\theta_i, \phi_j) : i = 1, ..., N_\theta;\ j = 1, ..., N_\phi\}$. The associated harmonic space representation is given by

$$\psi_m^\ell = \sum_{i,j} w_{ij} N_m^\ell P_m^\ell(\cos\theta_i) e^{-im\phi_j} \tag{38}$$

$$= \sum_i N_m^\ell P_m^\ell(\cos\theta_i) \sum_j w_{ij} e^{-im\phi_j}, \tag{39}$$

where fast Fourier transforms may be leveraged to compute the inner sum if the Dirac deltas are spaced evenly azimuthally (e.g. if $\phi_j = 2\pi j / N_\phi$). Alternative arbitrary samplings can of course be considered if useful for a problem at hand.

When defining filters in this manner one should be careful not to over-parametrize by assigning more weights than needed to define a filter at the harmonic bandlimit of the signal with which we wish to convolve. For example, if the filter is to be convolved with a signal bandlimited at L then a maximum of $2L - 1$ Dirac deltas should be placed along each ring of constant θ. One may also choose to interpolate the weights from a smaller number of learnable parameters acting as anchor points, allowing higher resolution filters to be defined with fewer learnable parameters.

C.2 Dirac Delta Filters on the Rotation Group

Similarly a Dirac delta function $\delta_{\rho'}$ on the rotation group $\mathrm{SO}(3)$ centered at position $\rho' = (\alpha', \beta', \gamma') \in \mathrm{SO}(3)$ is defined as

$$\delta_{\rho'}(\rho) = \frac{1}{\sin\beta}\delta_{\mathbb{R}}(\alpha - \alpha')\delta_{\mathbb{R}}(\cos\beta - \cos\beta')\delta_{\mathbb{R}}(\gamma - \gamma'), \tag{40}$$

with harmonic form

$$(\delta_{\rho'})_{mn}^\ell = D_{mn}^\ell(\rho') = e^{-im\alpha'} d_{mn}^\ell(\beta') e^{-in\gamma'}, \tag{41}$$

where d_{mn}^ℓ are Wigner (small) d-matrices.

The filter $\psi \in L^2(\mathrm{SO}(3))$ corresponding to a weighted sum of Dirac deltas with weights w_{ijk} assigned to Dirac delta functions centered at points $\{(\alpha_i, \beta_j, \gamma_k) : i = 1, ..., N_\alpha;\ j = 1, ..., N_\beta;\ k = 1, ..., N_\gamma\}$ has harmonic form

$$\psi_{mn}^\ell = \sum_{i,j,k} w_{ijk}\, e^{-im\alpha_j} d_{mn}^\ell(\beta_i) e^{-in\gamma_k} \tag{42}$$

$$= \sum_j d_{mn}^\ell(\beta_i) \sum_i e^{-im\alpha_j} \sum_k w_{ijk} e^{-in\gamma_k}, \tag{43}$$

where again fast Fourier transforms may be leveraged to compute the inner two sums assuming the Dirac deltas are spaced evenly in α and γ. The outer sums of Equation 39 and Equation 43 can also be computed by fast Fourier transforms by decomposing the Wigner d-matrices into their Fourier representation (cf. Trapani & Navaza, 2006; McEwen & Wiaux, 2011). One should again be careful not to over-parametrize.

D Equivariance Tests

To test rotational equivariance of operators we consider $N_f = 100$ random signals $\{f_i\}_{i=1}^{N_f}$ in $L^2(\Omega_1)$ with harmonic coefficients sampled from the standard normal distribution and $N_\rho = 100$

random rotations $\{\rho_j\}_{j=1}^{N_\rho}$ sampled uniformly on $\mathrm{SO}(3)$. In order to measure the extent to which an operator $\mathcal{A} : L^2(\Omega_1) \to L^2(\Omega_2)$ is equivariant we evaluate the mean relative error

$$d(\mathcal{A}(\mathcal{R}_{\rho_j} f_i), \mathcal{R}_{\rho_j}(\mathcal{A} f_i)) = \frac{1}{N_f} \frac{1}{N_\rho} \sum_{i=1}^{N_f} \sum_{j=1}^{N_\rho} \frac{\|\mathcal{A}(\mathcal{R}_{\rho_j} f_i) - \mathcal{R}_{\rho_j}(\mathcal{A} f_i))\|}{\|\mathcal{A}(\mathcal{R}_{\rho_j} f_i)\|} \tag{44}$$

resulting from pre-rotation of the signal, followed by application of $\mathcal{A}$, as opposed to post-rotation after application of $\mathcal{A}$, where the operator norm $\|\cdot\|$ is defined using the inner product $\langle \cdot, \cdot \rangle_{L^2(\Omega_2)}$.

Table 4 presents the mean relative equivariance errors computed. We consider the three standard convolutions described in Appendix B (with a random filter ψ_i for each signal f_i, generated in the same manner as f_i), the pointwise ReLU activation described in Section 2.5.1 for signals on the sphere ($\Omega_1 = \mathbb{S}^2$) and rotation group ($\Omega_1 = \mathrm{SO}(3)$), and the composition of tensor-product activation with a generalized convolution, described in Sections 2.5.2 and 2.4, respectively. We follow the tensor-product activation with a generalized convolution in order to project down onto the sphere to allow the same notion of error to be adopted as for the other operators. For consistency with the context in which we leverage these operators, all experiments are performed using single-precision arithmetic.

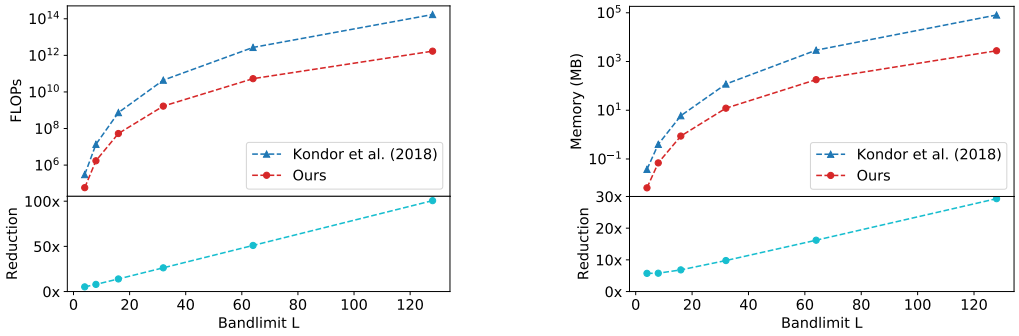
We see that all three standard notions of convolution and the composition of the tensor-product activation and generalized convolution are all strictly equivariant to floating point machine precision, with errors on the order of 10^{-7}. The pointwise ReLU operator is not strictly equivariant, with a mean relative error of 0.37 for signals on the rotation group and 0.34 for signals on the sphere. These errors reduce when the signals are oversampled before application of the ReLU, indicating that the error is due to aliasing induced by the spreading of information to higher degrees not captured at the original bandlimit. For example, for the pointwise ReLU operator on the rotation group oversampling by factors of $2\times, 4\times$ and $8\times$ results in a reduction in the mean relative equivariance error from 0.37 to $0.098, 0.032$ and 0.0096, respectively.

Table 4: Layer equivariance tests

Layer	Mean Relative Error
$\mathbb{S}^2$ to $\mathbb{S}^2$ conv.	4.4×10^{-7}
$\mathbb{S}^2$ to $\mathrm{SO}(3)$ conv.	5.3×10^{-7}
$\mathrm{SO}(3)$ to $\mathrm{SO}(3)$ conv.	9.3×10^{-7}
Tensor-product activation → Generalized conv.	5.0×10^{-7}
$\mathbb{S}^2$ ReLU	3.4×10^{-1}
$\mathbb{S}^2$ ReLU ($2\times$ oversampling)	8.9×10^{-2}
$\mathbb{S}^2$ ReLU ($4\times$ oversampling)	2.9×10^{-2}
$\mathbb{S}^2$ ReLU ($8\times$ oversampling)	1.3×10^{-2}
$\mathrm{SO}(3)$ ReLU	3.7×10^{-1}
$\mathrm{SO}(3)$ ReLU ($2\times$ oversampling)	9.8×10^{-2}
$\mathrm{SO}(3)$ ReLU ($4\times$ oversampling)	3.2×10^{-2}
$\mathrm{SO}(3)$ ReLU ($8\times$ oversampling)	9.6×10^{-3}

E Connection Between the Tensor Product Activation and Pointwise Squaring

To provide some intuition on the manner in which the tensor-product based activation introduces non-linearity into representations we describe its relationship to pointwise squaring for signals on the sphere. Here we consider the operator $\mathcal{N} : L^2(\mathbb{S}^2) \to L^2(\mathbb{S}^2)$ satisfying $(\mathcal{N} f)(x) = f^2(x)$ for all $x \in \mathbb{S}^2$, which differs subtly to $\mathcal{N}_\sigma$ with $\sigma(x) = x^2$ (using notation from Section 2.5.1), which corresponds to obtaining a sample-based representation at a finite bandlimit L and applying the squaring at the sample positions. For the special case $\sigma(x) = x^2$ we can directly compute the harmonic representation corresponding to the equivariance-preserving continuous limit $L \to \infty$.



(a) Computational cost

(b) Memory requirements

Figure 3: Comparison of computation and memory footprints between the generalized spherical CNN layers of Kondor et al. (2018) and our efficient generalized layers. The reduction in cost due to our efficient layers is given as multiplicative factors in the lower plot of each panel.

Given a spherical signal $f \in L^2(\mathbb{S}^2)$ with generalized representation $f = \{\hat{f}_0^\ell \in \mathbb{C}^{2\ell+1} : \ell = 0, .., L-1\}$, the generalized representation of the signal $f^2 \in L^2(\mathbb{S}^2)$ is given as

$$f^2 = \{ \sum_{(\ell_1,\ell_2)\in\mathcal{P}_{\ell,L}} (G^{\ell_1,\ell_2,\ell})^\top (\hat{f}_0^{\ell_1} \otimes \hat{f}_0^{\ell_2}) \ : \ \ell = 0, 1, ..., L-1\}, \tag{45}$$

where $G^{\ell_1,\ell_2,\ell} \in \mathbb{C}^{(2\ell_1+1)\times(2\ell_2+1)\times(2\ell+1)}$ are Gaunt coefficients defined as

$$G^{j_1 j_2 j_3}_{m_1 m_2 m_3} = \int_{\mathbb{S}^2} d\mu(\omega) Y^{j_1}_{m_1}(\omega) Y^{j_2}_{m_2}(\omega) Y^{j_3 *}_{m_3}(\omega). \tag{46}$$

Gaunt coefficients are related to the Clebsch-Gordan coefficients by

$$G^{j_1 j_2 j_3}_{m_1 m_2 m_3} = w^{j_1 j_2 j_3} C^{j_1 j_2 j_3}_{m_1 m_2 m_3}, \tag{47}$$

where $w^{j_1 j_2 j_3} = (-1)^{m_3} \sqrt{\frac{(2j_1+1)(2j_2+1)}{4\pi(2j_3+1)}} C^{j_1 j_2 j_3}_{0\,0\,0}$. Therefore, the continuous squaring operation corresponds to passing f through a tensor-product activation $\mathcal{N}_\otimes$ followed by a generalized convolution back down onto the sphere (single fragment per degree) with weight assigned to the $(\ell_1, \ell_2) \in \mathbb{P}_L^\ell$ fragment in degree-ℓ given by $w^{l_1 l_2 l_3}$.

This demonstrates that activations that are learnable within our framework can have very simple real-space interpretations. Even when confining outputs to the sphere we found it to be beneficial to allow the down-projection to be learnable rather than enforcing the weights given above for pointwise squaring. Learned activations will remain quadratic, however, given that output fragments are linear combinations of products between input fragments.

F Comparison of Computational and Memory Footprints

We perform a quantitative analysis of the computational cost and memory requirements of our proposed layers, and comparisons to prior approaches, to demonstrate how the complexity savings made through our proposals in Section 3 translate to tangible efficiency improvements.

We consider the simplest comparison, between a multi-channel generalized signal $f = (f_1, ..., f_K) \in \boldsymbol{\mathcal{F}}_L^K$ where each channel f_i has type $\tau_{f_i} = (1, ..., 1)$ (and therefore corresponds to a signal on the sphere) and a uni-channel signal $g \in \boldsymbol{\mathcal{F}}_L$ of type $\tau^g = (K, ..., K)$. The setting corresponding to signal f captures the efficiency improvements of our proposed layers, while the setting corresponding to g represents the case without these improvements. Notice that the total number of fragments is the same in both f and g. We compare the number of floating point operations (FLOPs) and amount of memory required to perform a tensor-product based activation followed by a generalized convolution projecting back down onto a signal of the same type as the input. When applied to f, MST mixing sets $\widetilde{\mathbb{P}}_L^\ell$ (Section 3.1.3) and constrained generalized convolutions (Section 3.1.2) are

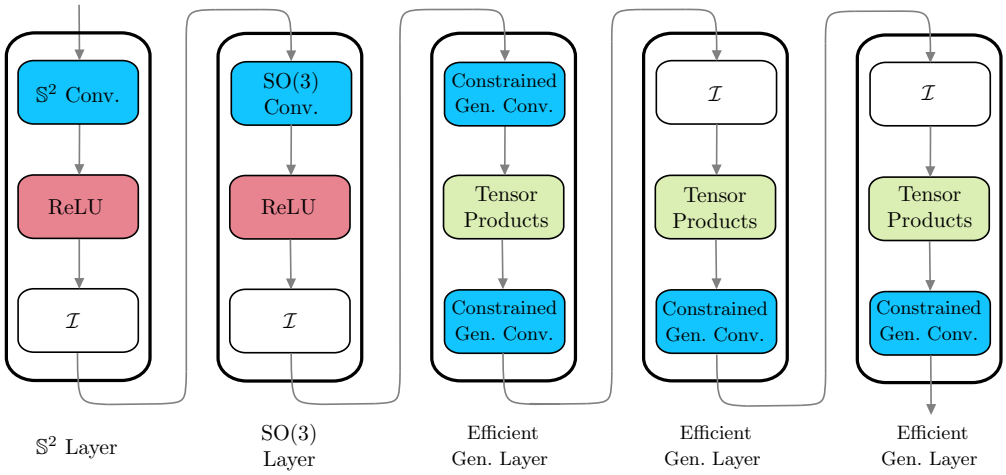


Figure 4: Visualization of the architecture used for the convolutional base in our hybrid models. The input to the first convolutional layer is a signal on the sphere. The output from the final convolutional layer are scalar values corresponding to fragments of degree $\ell = 0$, which are then mapped through some fully connected layers to give the model output.

used. When applied to g full mixing sets $\mathbb{P}_L^\ell$ and unconstrained generalized convolutions are used, as in Kondor et al. (2018). Considered are the costs for a single training instance (batch size of 1).

Figure 3 shows the computational costs and memory requirements, in terms of floating point operations and megabytes respectively, for $K = 4$ and various spatial bandlimits L. We adopt the convention whereby complex addition and multiplication require 2 and 6 floating point operations respectively. At low bandlimits we see the saving arising from the channel-wise structure. Note that the saving illustrated here is relatively small since $K = 4$ for these experiments (so that the L scaling is apparent), whereas in practice typically $K \sim 100$. The saving then increases linearly in response to increases in the bandlimit of the input, as expected given the $\mathcal{O}(L^5)$ and $\mathcal{O}(L^4)$ spatial complexities. We see that even in this simple case, with a relatively small number of channels ($K = 4$), both the computational and memory footprints are reduced by orders of magnitude. At a bandlimit of $L = 128$ the computational cost is 101-times reduced and the memory requirement is 29-times reduced.

G Additional Information on Experiments

G.1 Rotated MNIST on the Sphere

For our MNIST experiments we used a hybrid model with the architecture shown in Figure 4. The first block includes a directional convolution on the sphere that lifts the spherical input ($\tau^\ell_{f^{(0)}} = 1$) onto the rotation group ($\tau^\ell_{f^{(1)}} = \min(2\ell + 1, 2N_1 - 1)$). The second block includes a convolution on the rotation group, hence its input and output both live on the rotation group. We then apply a restricted generalized convolution to map to type $\tau^\ell_{f^{(3)}} = \lceil \tau_{\text{max}}/\sqrt{2\ell + 1} \rceil$, where τ_{max}=5. The same type is used for the following three channel-wise tensor-product activations and two restricted generalized convolutions until the final restricted generalized convolution maps down to a rotationally invariant representation ($\tau^\ell_{f^{(5)}} = \delta_{\ell 0}$). As is traditional in convolution networks we gradually decrease the resolution, with $(L_0, L_1, L_2, L_3, L_4, L_5) = (20, 10, 10, 6, 3, 1)$, and increase the number of channels, with $(K_0, K_1, K_2, K_3, K_4, K_5) = (1, 20, 22, 24, 26, 28)$. We proceed these convolutional layers with a single dense layer of size 30, sandwiched between two dropout layers (keep probability 0.5), and then fully connect to the output of size 10.

We train the network for 50 epochs on batches of size 32, using the Adam optimizer (Kingma & Ba, 2015) with a decaying learning rate starting at 0.001. For the restricted generalized convolutions we follow the approach of Kondor et al. (2018) by using L_1 regularization (regularization strength 10^{-5}) and applying a restricted batch normalization across fragments, where the fragments are only scaled by their average and not translated (to preserve equivariance).

G.2 ATOMIZATION ENERGY PREDICTION

When regressing the atomization energy of molecules there are two inputs to the model: the number of atoms of each element contained in the molecule; and spherical cross-sections of the potential energy around each atom. We adopt the high-level QM7-specific architecture of Cohen et al. (2018) which contains a spherical CNN as a sub-model, for which we substitute our own. This results in an overall model that is invariant to both rotations of the molecule around each constituent atom and to permutations of the ordering of the atoms.

The first (non-spherical) input is mapped onto a scalar output using a multi-layer perceptron (MLP) with three hidden layers of sizes 100, 100 and 10 (and ReLU activations). The second input, multiple spherical cross sections for each atom, are separately projected using a shared spherical CNN (of architecture described below) onto lower dimensional vectors of size 64. The mean vector is then taken across atoms (ensuring invariance w.r.t. permutations of the atoms) and mapped onto a scalar output using an MLP with a single hidden layer of size 512 (with a ReLU activation). The predicted energy is then taken to be the sum of the two scalar outputs.

As a starting point we train the first MLP to regress the atomization energies alone (achieving RMS $\sim$20), before pairing it with the spherical model (and its connected MLP). We then train the joint model for 60 epochs, again with the Adam optimizer, a decaying learning rate (starting at 2.5×10^{-4}), regularizing the efficient generalized layers with L_2 regularization (strength 2.5×10^{-6}) and batch sizes of 32.

For the spherical component we again adopt the convolutional architecture shown in Figure 4 except with one fewer efficient generalized layer. We use bandlimits of $(L_0, L_1, L_2, L_3, L_4)$=$(10, 6, 6, 3, 1)$, channels of $(K_0, K_1, K_2, K_3, K_4)$=$(5, 16, 24, 32, 40)$ and $\tau_{\text{max}} = 6$. One minor difference is that this time we include a skip connection between the $\ell = 0$ components of the fourth and fifth layer. We proceed the convolutional layers with two dense layers of size $(256, 64)$ and use batch normalization between each layer.

G.3 3D SHAPE RETRIEVAL

To project the 3D meshes of the SHREC'17 data onto spherical representations (bandlimited at $L = 128$) we adopt the preprocessing approach of Cohen et al. (2018) and augment the data with random rotations and translations.

We construct a model with an architecture that is again similar to that described in Appendix G.1 but with an additional axisymmetric convolutional layer prepended to the start of the network and one fewer efficient generalized layers. We use bandlimits $(L_0, L_1, L_2, L_3, L_4, L_5) = (128, 32, 16, 16, 6, 1)$, channels $(K_0, K_1, K_2, K_3, K_4, K_5) = (6, 20, 30, 40, 60, 70)$ and $\tau_{\text{max}} = 6$ for the efficient generalized layers. The convolutional layers are followed by a dense layer of size 128 which is fully connected to the output (of size 55).

We again train with the Adam optimizer, a decaying learning rate (starting at 5×10^{-4}) and batch sizes of 8, this time until performance on the validation set showed no improvement for at least 4 epochs (36 epochs in total). We perform batch normalization between convolutional layers and dropout preceding the dense layer. We regularize the efficient generalized layers with L_2 regularization (strength 10^{-5}). When testing our model we average the output probabilities over 15 augmentations of the data.