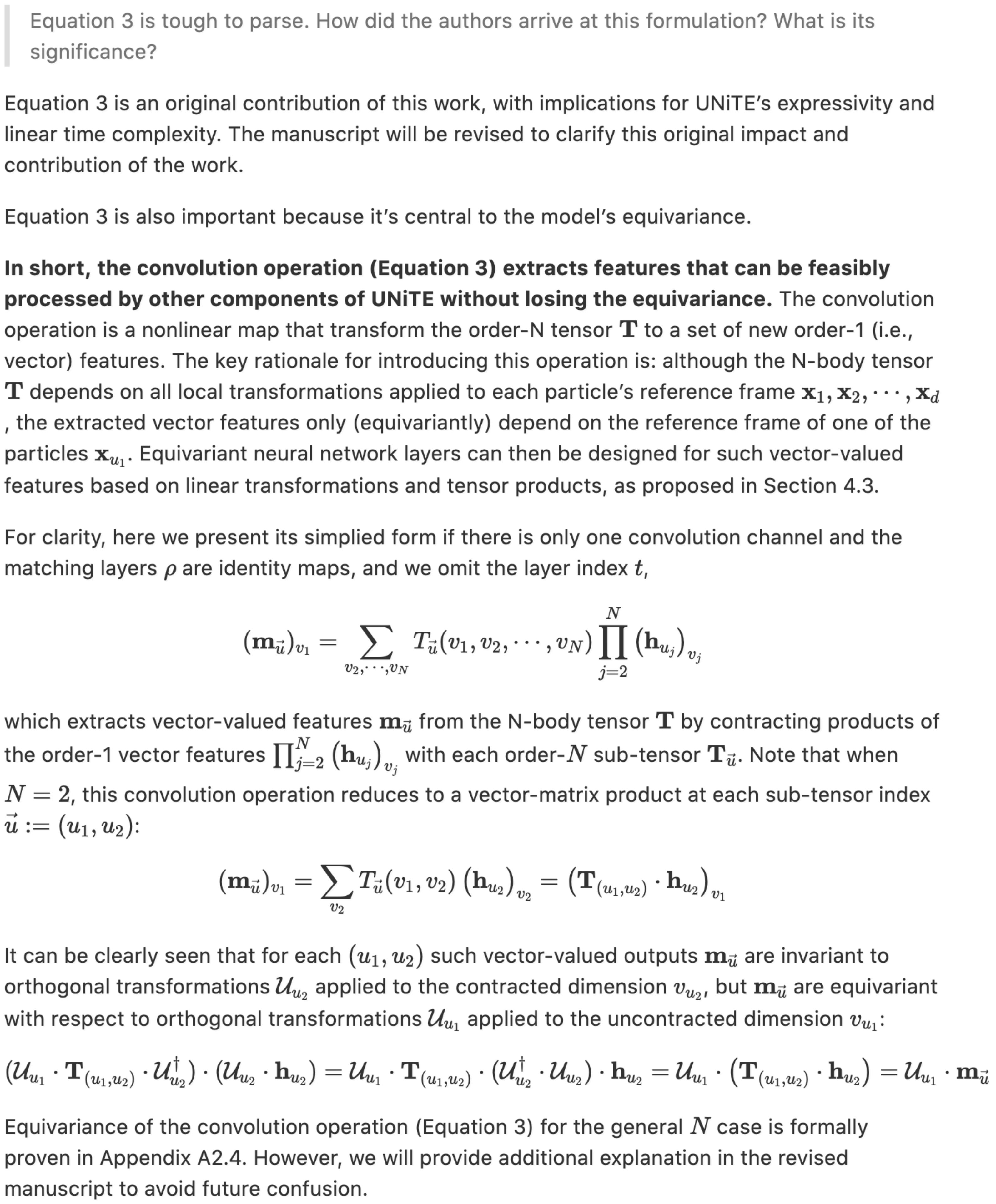

> Equation 3 is tough to parse. How did the authors arrive at this formulation? What is its significance?

Equation 3 is an original contribution of this work, with implications for UNiTE's expressivity and linear time complexity. The manuscript will be revised to clarify this original impact and contribution of the work.

Equation 3 is also important because it's central to the model's equivariance.

In short, the convolution operation (Equation 3) extracts features that can be feasibly processed by other components of UNiTE without losing the equivariance. The convolution operation is a nonlinear map that transform the order-N tensor $\mathbf{T}$ to a set of new order-1 (i.e., vector) features. The key rationale for introducing this operation is: although the N-body tensor $\mathbf{T}$ depends on all local transformations applied to each particle's reference frame $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_d$, the extracted vector features only (equivariantly) depend on the reference frame of one of the particles $\mathbf{x}_{u_1}$. Equivariant neural network layers can then be designed for such vector-valued features based on linear transformations and tensor products, as proposed in Section 4.3.

For clarity, here we present its simplied form if there is only one convolution channel and the matching layers ρ are identity maps, and we omit the layer index t,

$$(\mathbf{m}_{\vec{u}})_{v_1} = \sum_{v_2,\cdots,v_N} T_{\vec{u}}(v_1, v_2, \cdots, v_N) \prod_{j=2}^{N} \left(\mathbf{h}_{u_j}\right)_{v_j}$$

which extracts vector-valued features $\mathbf{m}_{\vec{u}}$ from the N-body tensor $\mathbf{T}$ by contracting products of the order-1 vector features $\prod_{j=2}^{N} \left(\mathbf{h}_{u_j}\right)_{v_j}$ with each order-N sub-tensor $\mathbf{T}_{\vec{u}}$. Note that when $N = 2$, this convolution operation reduces to a vector-matrix product at each sub-tensor index $\vec{u} := (u_1, u_2)$:

$$(\mathbf{m}_{\vec{u}})_{v_1} = \sum_{v_2} T_{\vec{u}}(v_1, v_2) \left(\mathbf{h}_{u_2}\right)_{v_2} = \left(\mathbf{T}_{(u_1,u_2)} \cdot \mathbf{h}_{u_2}\right)_{v_1}$$

It can be clearly seen that for each (u_1, u_2) such vector-valued outputs $\mathbf{m}_{\vec{u}}$ are invariant to orthogonal transformations $\mathcal{U}_{u_2}$ applied to the contracted dimension v_{u_2}, but $\mathbf{m}_{\vec{u}}$ are equivariant with respect to orthogonal transformations $\mathcal{U}_{u_1}$ applied to the uncontracted dimension v_{u_1}:

$$(\mathcal{U}_{u_1} \cdot \mathbf{T}_{(u_1,u_2)} \cdot \mathcal{U}_{u_2}^{\dagger}) \cdot (\mathcal{U}_{u_2} \cdot \mathbf{h}_{u_2}) = \mathcal{U}_{u_1} \cdot \mathbf{T}_{(u_1,u_2)} \cdot (\mathcal{U}_{u_2}^{\dagger} \cdot \mathcal{U}_{u_2}) \cdot \mathbf{h}_{u_2} = \mathcal{U}_{u_1} \cdot \left(\mathbf{T}_{(u_1,u_2)} \cdot \mathbf{h}_{u_2}\right) = \mathcal{U}_{u_1} \cdot \mathbf{m}_{\vec{u}}$$

Equivariance of the convolution operation (Equation 3) for the general N case is formally proven in Appendix A2.4. However, we will provide additional explanation in the revised manuscript to avoid future confusion.