Invariance and Equivariance in IPA: Local vs. Global Frames

Study Notes

December 8, 2025

1 Introduction

This note rebuilds the picture of local vs. global coordinates, rigid transforms, and the invariant point attention (IPA) mechanism from the ground up. The aims are:

1. Explain what *local* and *global* coordinates mean in very concrete terms.
2. Show how the transforms T_i work and why the distances in the IPA logits are invariant under a global rigid motion of the protein.
3. Walk through the three outputs of IPA, focusing on the third one where T_i^{-1} and T_j appear together.

Notation is aligned with Algorithm 22 in the AF2 supplement.

2 Local vs. Global Coordinates

2.1 Intuitive Picture

Imagine you are in a room.

- On the floor there is a large grid: x-axis left–right, y-axis front–back, z-axis up. Distances are in meters. This is the *global coordinate system* of the room.
- You stand somewhere in the room and define your own private axes:
  - x-axis: from your chest to your right hand,
  - y-axis: from your chest to your head,
  - z-axis: from your chest backward through your spine.

  This is your *local coordinate system*, with origin at your chest.

Pick the tip of your right index finger.

- In room coordinates, its coordinates might be $(x, y, z) = (2.3, 1.1, 1.5)$,
- In your local coordinates, it might be (0.6, 0.0, 0.0), meaning "60 cm along my right arm, not up, not backward".

This is the same physical point, described in two coordinate systems.

Now imagine you rotate your whole body and walk to another place in the room.

- The room coordinates of your finger change substantially,
- The local coordinates of your finger relative to your chest hardly change, because your pose is the same.

This is essentially the idea of a local frame vs. a global frame in AlphaFold.

- **Global frame:** one fixed 3D coordinate system covering the whole protein,
- **Local frame of residue i:** origin at its Cα atom, axes aligned to its backbone geometry.

Local coordinates describe where something is relative to a particular residue; global coordinates describe where it is in the protein as a whole.

3 Rigid Transforms and Frames: Mathematical Formulation

3.1 Rigid Transforms

A rigid transform T consists of

- a rotation matrix $R \in \mathbb{R}^{3\times 3}$,
- a translation vector $t \in \mathbb{R}^3$,

and we write $T = (R, t)$.

Given a point $x \in \mathbb{R}^3$, the transform acts as

$$T \circ x = Rx + t.$$

Rotations preserve lengths and angles; translations shift all points by the same amount without changing distances. Any combination of rotation and translation is a rigid motion preserving Euclidean distances.

3.2 Frames for Residues

A frame for residue i is a rigid transform

$$T_i = (R_i, t_i)$$

with the interpretation:

- the local frame has origin at the Cα of residue i,
- the axes are determined by the N, Cα, C atoms via a Gram–Schmidt construction.

Then:

- A local coordinate $u_i \in \mathbb{R}^3$ describes a point in residue i's local frame,
- The corresponding global coordinate is

$$x = T_i \circ u_i = R_i u_i + t_i.$$

In words, "apply T_i" means: go from residue-i local coordinates to global 3D coordinates.

3.3 Inverse Transform T_i^{-1}

If $T = (R, t)$, the inverse transform T^{-1} is defined so that

$$T^{-1} \circ (T \circ x) = x$$

for all points x.

Since $T \circ x = Rx + t$, solving for x in terms of $y = T \circ x$ gives

$$x = R^{-1}(y - t).$$

Thus

$$T^{-1} = (R^{-1}, -R^{-1}t)$$

and

$$T^{-1} \circ y = R^{-1}(y - t).$$

Because R is a rotation, $R^{-1} = R^\top$.

Conceptually:

- T takes you from local to global coordinates,
- T^{-1} takes you from global back to local.

This is why expressions such as $T_j \circ \tilde{v}_j^{hp}$ (local to global) and $T_i^{-1} \circ (\dots)$ (global back to local) appear in IPA.

4 Point Features: $\tilde{q}$, $\tilde{k}$, $\tilde{v}$

In Algorithm 22, for each residue i and head h, the network predicts:

- query vectors $q_i^h \in \mathbb{R}^c$ and key vectors $k_j^h \in \mathbb{R}^c$ (as in a standard Transformer),
- query points $\tilde{q}_i^{hp} \in \mathbb{R}^3$ for $p = 1, \dots, N_{\text{query points}}$ in the local frame of residue i,
- key points $\tilde{k}_j^{hp} \in \mathbb{R}^3$ in the local frame of residue j,
- value points $\tilde{v}_j^{hp} \in \mathbb{R}^3$ in the local frame of residue j.

Each head therefore sees a small constellation of points attached to each residue, expressed as offsets from that residue's Cα in its own local coordinates.

5 The Geometric Term in the Attention Logit

We now rewrite the attention logit from Algorithm 22, line 7 in a compact way. For each head h, query residue i, and key residue j:

$$a_{ij}^h = \text{softmax}_j\left(w_L\left[\sqrt{\frac{1}{c}}(q_i^h)^\top k_j^h + b_{ij}^h - \gamma_h w_C^2 \sum_p \left\| T_i \circ \tilde{q}_i^{hp} - T_j \circ \tilde{k}_j^{hp} \right\|^2\right]\right),$$

where:

- $q_i^h, k_j^h \in \mathbb{R}^c$ are the usual query and key vectors,
- b_{ij}^h is a scalar pair bias from the pair representation z_{ij},
- γ_h is a positive learned scalar per head,
- w_C, w_L are fixed scaling constants chosen via a variance calculation.

The geometric part is

$$D_{ij}^h = \sum_p \left\| T_i \circ \tilde{q}_i^{hp} - T_j \circ \tilde{k}_j^{hp} \right\|^2.$$

5.1 One Point Index p

Fix i, j, h and a particular point index p.

1. $\tilde{q}_i^{hp} \in \mathbb{R}^3$ is in local coordinates of residue i; think of it as a direction and distance from the Cα of i in its own coordinate system.
2. $T_i = (R_i, t_i)$ is the frame of residue i. Here t_i is the Cα position of i in global space and R_i aligns the local axes to global axes.
3. Thus $T_i \circ \tilde{q}_i^{hp} = R_i \tilde{q}_i^{hp} + t_i$ is a 3D point in global coordinates.
4. Similarly, $\tilde{k}_j^{hp}$ is in residue j's local frame and $T_j \circ \tilde{k}_j^{hp} = R_j \tilde{k}_j^{hp} + t_j$ is its global counterpart.

Now both points live in the same global coordinate system, so we can subtract them and take a distance:

$$d_{ij}^{hp} = \left\| T_i \circ \tilde{q}_i^{hp} - T_j \circ \tilde{k}_j^{hp} \right\|.$$

The geometric term sums over all p:

$$D_{ij}^h = \sum_p \left(d_{ij}^{hp} \right)^2.$$

The logit subtracts $\gamma_h w_C^2 D_{ij}^h$, so large distances between the point constellations for i and j reduce the logit and hence the attention weight.

In words:

> Head h prefers residue i to attend to residues j whose learned key points lie close in 3D to the learned query points of i, when interpreted through the current backbone frames.

This is why the authors call IPA "invariant point attention": it is attention that depends on 3D point positions, not only feature similarity.

6 Invariance Under Global Rotations and Translations

We now show why the geometric term, and therefore the logits, are invariant to a global rigid motion of the protein.

6.1 Defining the Global Move

Suppose we choose any global rigid transform $T_{\text{global}} = (R_g, t_g)$. Define new frames

$$T_i' = T_{\text{global}} \circ T_i.$$

By composition of transforms,

$$T_i' \circ x = T_{\text{global}} \circ (T_i \circ x) = R_g(R_i x + t_i) + t_g.$$

Geometrically, we have:

- Rotated and translated the entire structure,
- Left all local coordinates $\tilde{q}_i^{hp}$, $\tilde{k}_j^{hp}$ unchanged.

This is akin to picking up the protein and moving it rigidly in space.

6.2 Effect on a Single Distance Term

Consider a single squared distance in the geometric term, but under the new frames T_i':

$$\left\| T_i' \circ \tilde{q}_i^{hp} - T_j' \circ \tilde{k}_j^{hp} \right\|^2.$$

By definition of T_i',

$$T_i' \circ \tilde{q}_i^{hp} = T_{\text{global}} \circ T_i \circ \tilde{q}_i^{hp},$$

and similarly for j. Thus the distance becomes

$$\left\| T_{\text{global}} \circ T_i \circ \tilde{q}_i^{hp} - T_{\text{global}} \circ T_j \circ \tilde{k}_j^{hp} \right\|^2.$$

Rigid transforms preserve distances: for any $x, y \in \mathbb{R}^3$,

$$\left\| T_{\text{global}} \circ x - T_{\text{global}} \circ y \right\|^2 = \|x - y\|^2.$$

To see this, write $T_{\text{global}} \circ x = R_g x + t_g$ and $T_{\text{global}} \circ y = R_g y + t_g$; then

$$\begin{aligned} \left\| T_{\text{global}} \circ x - T_{\text{global}} \circ y \right\|^2 &= \left\| R_g x + t_g - (R_g y + t_g) \right\|^2 \\ &= \left\| R_g(x - y) \right\|^2 \\ &= (x - y)^\top R_g^\top R_g (x - y) \\ &= (x - y)^\top I (x - y) \\ &= \|x - y\|^2, \end{aligned}$$

since R_g is a rotation and $R_g^\top R_g = I$.

Applying this with

$$x = T_i \circ \tilde{q}_i^{hp}, \qquad y = T_j \circ \tilde{k}_j^{hp},$$

we obtain

$$\left\| T_{\text{global}} \circ T_i \circ \tilde{q}_i^{hp} - T_{\text{global}} \circ T_j \circ \tilde{k}_j^{hp} \right\|^2 = \left\| T_i \circ \tilde{q}_i^{hp} - T_j \circ \tilde{k}_j^{hp} \right\|^2.$$

Thus each squared distance is unchanged. Summing over p,

$$D_{ij}^{h,\prime} = \sum_p \left\| T_i' \circ \tilde{q}_i^{hp} - T_j' \circ \tilde{k}_j^{hp} \right\|^2 = \sum_p \left\| T_i \circ \tilde{q}_i^{hp} - T_j \circ \tilde{k}_j^{hp} \right\|^2 = D_{ij}^h.$$

6.3 Invariance of the Full Logit

Revisit the logit inside the softmax:

$$\text{logit}_{ij}^h = w_L \Big[\sqrt{\tfrac{1}{c}} (q_i^h)^\top k_j^h + b_{ij}^h - \gamma_h w_C^2 D_{ij}^h \Big].$$

Under $T_i \mapsto T_i'$:

- The feature terms q_i^h, k_j^h, b_{ij}^h do not depend on frames,
- We have just shown $D_{ij}^{h,\prime} = D_{ij}^h$.

Therefore logit_{ij}^h does not change. Feeding the same logits into softmax_j yields identical attention weights a_{ij}^h.

Hence the weights are invariant to any global rigid motion, capturing the idea that IPA depends only on relative geometry.

7 Outputs of IPA and Global-to-Local Mapping

Once we have the attention weights a_{ij}^h, Algorithm 22 computes three outputs. The third is where the combination $T_j \circ \tilde{v}_j^{hp}$ and T_i^{-1} appears.

7.1 First Output: Pair-Weighted

Line 8:

$$\tilde{o}_i^h = \sum_j a_{ij}^h z_{ij}.$$

Here z_{ij} is the pair representation between residues i and j; for each head h, $\tilde{o}_i^h$ aggregates pair features of residues that i attends to.

7.2 Second Output: Standard Value Aggregation

Line 9:

$$o_i^h = \sum_j a_{ij}^h v_j^h.$$

This is standard attention over value vectors v_j^h in feature space; no coordinates appear here.

7.3 Third Output: Value Points and Local Mapping

The third output, line 10:

$$\tilde{o}_i^{hp} = T_i^{-1} \circ \Big(\sum_j a_{ij}^h \big(T_j \circ \tilde{v}_j^{hp} \big) \Big),$$

can be decomposed as follows.

1. Local to global for value points. For each residue j and value point index p, we have $\tilde{v}_j^{hp}$ in j's local frame. Global coordinates are

$$g_j^{hp} = T_j \circ \tilde{v}_j^{hp} = R_j \tilde{v}_j^{hp} + t_j.$$

2. Attention-weighted barycenter in global space. For residue i, head h, and point p,

$$y_i^{hp} = \sum_j a_{ij}^h g_j^{hp} = \sum_j a_{ij}^h \big(T_j \circ \tilde{v}_j^{hp}\big).$$

This is the attention-weighted average of neighbor points in global 3D, i.e. a barycenter.

3. Global back to local for residue i. Apply residue i's inverse frame:

$$\tilde{o}_i^{hp} = T_i^{-1} \circ y_i^{hp}.$$

This converts the global point into i's local coordinates.

Thus $\tilde{o}_i^{hp} \in \mathbb{R}^3$ encodes "where is the average of neighbors' value points, as seen from residue i's coordinate system".

7.4 Why Map Back to Local Coordinates?

If we kept y_i^{hp} in global coordinates and fed it directly forward, then:

- Rotating the entire protein by a global transform would change these coordinates,
- Downstream MLPs would see different numbers for the same physical shape under different global orientations,
- The network would not be equivariant and would need heavy augmentation over random orientations.

By expressing coordinates back in residue-i's local frame, we obtain invariance:

- The shape of the local neighborhood around i is described in i's coordinate system, which moves with i,
- If we rotate and translate the protein, both i and its neighbors move, but their relative configuration in i's frame remains the same.

This matches the textual description in the AF2 manuscript:

> Each residue produces query points, key points and value points in its local frame. These points are projected into the global frame using the backbone frame of the residue in which they interact with each other. The resulting points are then projected back into the local frame. The coordinate transformations ensure the invariance of this module with respect to the global frame.

7.5 Invariance of the Value-Point Outputs

Under a global transform T_{global}, define

$$T_j' = T_{\text{global}} \circ T_j, \qquad T_i' = T_{\text{global}} \circ T_i.$$

First, transform each value point:

$$g_j^{hp} = T_j \circ \tilde{v}_j^{hp} \quad \Rightarrow \quad g_j^{hp,\prime} = T_j' \circ \tilde{v}_j^{hp} = T_{\text{global}} \circ T_j \circ \tilde{v}_j^{hp} = T_{\text{global}} \circ g_j^{hp}.$$

The weighted average in global space becomes

$$y_i^{hp,\prime} = \sum_j a_{ij}^h g_j^{hp,\prime} = \sum_j a_{ij}^h \big(T_{\text{global}} \circ g_j^{hp}\big) = T_{\text{global}} \circ \Big(\sum_j a_{ij}^h g_j^{hp}\Big) = T_{\text{global}} \circ y_i^{hp}.$$

Now compute the local output in the new frames:

$$\tilde{o}_i^{hp,\prime} = (T_i')^{-1} \circ y_i^{hp,\prime}.$$

Since $T_i' = T_{\text{global}} \circ T_i$, we have

$$(T_i')^{-1} = T_i^{-1} \circ T_{\text{global}}^{-1}.$$

Therefore

$$\tilde{o}_i^{hp,\prime} = T_i^{-1} \circ T_{\text{global}}^{-1} \circ \big(T_{\text{global}} \circ y_i^{hp}\big) = T_i^{-1} \circ y_i^{hp} = \tilde{o}_i^{hp}.$$

So the local point output $\tilde{o}_i^{hp}$ is invariant to global rigid motions, exactly as desired.

8 Symbol-by-Symbol Interpretation of a Typical Block

Consider the typical logit expression:

$$a_{ij}^h = \text{softmax}_j\Big(w_L\big[\sqrt{1/c}\,(q_i^h)^\top k_j^h + b_{ij}^h - \gamma_h w_C^2 \sum_p \|T_i \circ \tilde{q}_i^{hp} - T_j \circ \tilde{k}_j^{hp}\|^2\big]\Big).$$

Here:

- a_{ij}^h: attention weight from residue i to residue j in head h,
- softmax_j: softmax over the j index (weights for fixed i, h sum to 1),
- w_L: global scaling constant to keep logit variance under control at initialization,
- $\sqrt{1/c}(q_i^h)^\top k_j^h$: scaled dot product of queries and keys (standard content similarity),
- b_{ij}^h: learned bias from the pair representation,
- $-\gamma_h w_C^2 \sum_p \cdots$: penalty for geometric distance between point constellations after mapping them to global space.

Every term is either independent of frames (dot product, pair bias) or invariant to global rigid motion (distance term), so the whole logit is invariant.

9 A 2D Toy Example

A small two-dimensional example can make the invariance idea concrete.

Work in 2D for simplicity.

- Residue i has frame $T_i = (R_i, t_i)$ with R_i a 90° rotation and $t_i = (1, 0)$,
- Residue j has frame $T_j = (R_j, t_j)$ with R_j the identity and $t_j = (0, 1)$.

Suppose a single query and key point index p:

- $\tilde{q}_i^{hp} = (1, 0)$ in i's local frame,
- $\tilde{k}_j^{hp} = (0, 1)$ in j's local frame.

Local-to-global:

- $T_i \circ \tilde{q}_i^{hp} = R_i(1, 0) + (1, 0) = (0, 1) + (1, 0) = (1, 1)$,
- $T_j \circ \tilde{k}_j^{hp} = I(0, 1) + (0, 1) = (0, 2)$.

Distance term:

$$\left\| T_i \circ \tilde{q}_i^{hp} - T_j \circ \tilde{k}_j^{hp} \right\|^2 = \|(1, 1) - (0, 2)\|^2 = \|(1, -1)\|^2 = 1^2 + (-1)^2 = 2.$$

Now apply a global transform $T_{\text{global}} = (R_g, t_g)$ that rotates by 90° and translates by $(10, 0)$:

- $T_i' = T_{\text{global}} \circ T_i$,
- $T_j' = T_{\text{global}} \circ T_j$.

New global points:

- $T_i' \circ \tilde{q}_i^{hp} = T_{\text{global}} \circ (T_i \circ \tilde{q}_i^{hp}) = T_{\text{global}} \circ (1, 1) = R_g(1, 1) + (10, 0)$,
  rotating $(1, 1)$ by 90° gives $(-1, 1)$, so we get $(9, 1)$,
- $T_j' \circ \tilde{k}_j^{hp} = T_{\text{global}} \circ (0, 2) = R_g(0, 2) + (10, 0) = (-2, 0) + (10, 0) = (8, 0)$.

Distance:

$$\|(9, 1) - (8, 0)\|^2 = \|(1, 1)\|^2 = 1^2 + 1^2 = 2,$$

exactly the same as before. The geometric term therefore cannot distinguish whether we have moved and rotated the entire protein, which is exactly the intended invariance.

10 Big-Picture Intuition

Putting everything together:

- Each residue i has a frame T_i that mediates between local and global coordinates; local points are attached to the residue, global points are positions in 3D.
- IPA uses this bridge twice:

1. To compare query and key points from residues i and j in global space, biasing attention towards geometrically nearby residues,
2. To aggregate value points in global space and then re-express the averaged geometry in residue i's local frame.

- The composition $T_j \circ \tilde{v}_j^{hp}$ followed by T_i^{-1} is simply "local to global, aggregate, then global back to local".
- Since rigid transforms preserve distances and we always return to local frames at the end, everything that IPA produces is invariant to any global rotation or translation of the whole protein.

If any part of the local-to-global or global-to-local story remains unclear, one can always fall back on small algebraic examples or 2D cartoons with two residues and a single point, as sketched above.