Protein-Ligand Affinity as Multiscale Correspondence: A Takens-Based Programme for Sequence-to-Structure and Affinity Modelling - Part 2

Kevin R. Haylett
Manchester,UK

Selected Communications

May 24, 2026

The Construction Signal as a Literal Dynamical Time Series – Transcription, Introns, and Co-Translational Dynamics

Abstract

The original paper treats the amino-acid sequence as an observed one-dimensional symbolic construction signal whose hidden geometry can be reconstructed via Takens-style delay embeddings. This addendum clarifies and strengthens that framing: the biological construction process is not a metaphor but a literal dynamical time series. DNA is transcribed nucleotide-by-nucleotide into pre-mRNA; introns form an integral part of that temporal signal; splicing and translation proceed sequentially; and co-translational folding occurs while the chain is still being produced. Treating the full primary transcript (exons + introns) as the observable signal allows delay-coordinate methods to capture co-transcriptional and co-translational constraints directly. The mathematical structure, biological relationships, and practical modelling extensions are stated formally. A short research plan and open questions for future investigators are provided.

1 Purpose of the addendum

The central claim of the main paper–that affinity is a multiscale correspondence over compressed construction signals–remains unchanged. However, the phrase "ordered symbolic trace" (Section 2) can be sharpened. The residue sequence $P = (p_1, p_2, \ldots, p_N)$ is not an arbitrary static vector; it is the processed output of a genuine sequential dynamical process carried out by the cell's transcription and translation machinery. Introns are not extraneous; they are part of the raw time series that evolution has tuned for regulatory, structural, and evolvability purposes. This addendum makes the time-series nature explicit and shows how it fits naturally within the Takens-based programme.

2 Mathematical structure: extending delay embeddings to the primary transcript

Let the observable construction signal now be the full primary transcript sequence (pre-mRNA):

$$T = (t_1, t_2, \ldots, t_M) \tag{1}$$

where each t_j is a nucleotide (or codon-level token), $M \gg N$, and the signal includes both exons and introns. The mature protein sequence P is obtained after splicing:

$$P = \mathcal{S}(T) \tag{2}$$

where $\mathcal{S}$ denotes the (possibly alternative) splicing map.

A residue (or nucleotide) embedding is formed as before:

$$e_i = E(p_i) \quad \text{or} \quad e_j^T = E^T(t_j) \tag{3}$$

A delay-coordinate vector on the transcript signal is

$$\Phi_{\tau,d}^T(j) = \big(e_j^T,\ e_{j+\tau}^T,\ \ldots,\ e_{j+(d-1)\tau}^T\big). \tag{4}$$

A multiscale family of delays $T = \{\tau_1, \tau_2, \ldots, \tau_R\}$ produces the representation

$$\mathcal{E}(T) = \big\{\Phi_{\tau_r,d_r}^T(j) : r = 1, \ldots, R;\ j = 1, \ldots, M\big\}. \tag{5}$$

Short delays capture local splicing signals and codon usage; intermediate delays capture intron-mediated exon pairing and co-transcriptional RNA secondary structure; longer delays expose distant genomic positions that become spatially or functionally related after splicing and folding. The final protein geometry $\hat{X}_P$ and affinity prediction $\hat{a}$ are then produced from a joint representation that includes both the spliced protein delays $\mathcal{E}(P)$ and the upstream transcript delays $\mathcal{E}(T)$:

$$\hat{a} = H_\theta\big(\mathcal{E}(P),\ \mathcal{E}(T),\ \mathcal{E}_L(L),\ C_{PL}(P, L),\ q\big). \tag{6}$$

This is a direct extension of Equation (29) in the main paper. The splicing map $\mathcal{S}$ can itself be learned or treated as a differentiable operation within the model.

3 Biological relationships: the literal time-series construction process

- **Transcription** produces the pre-mRNA signal sequentially (RNA polymerase advances base-by-base in real time).
- **Introns** are retained in the raw transcript and carry essential regulatory information: splice-site signals, pausing elements, chromatin-looping anchors, and evolutionary modules for exon shuffling.
- **Splicing** is co-transcriptional or post-transcriptional but operates on the temporal signal T, not on a static bag of exons.
- **Translation** is strictly vectorial: the ribosome reads codons one-by-one, and the nascent polypeptide begins folding inside the exit tunnel while later residues are still being synthesized.

Consequently, long-range sequence dependencies in the final protein are shaped by temporal constraints that were already present in the primary transcript.

The affinity label a therefore supervises a correspondence that spans the entire dynamical history: from genomic DNA through the full transcript time series to the folded multiscale protein object.

4 Modelling implications and research directions

The literal time-series view offers four immediate practical extensions to the Takens-based Transformer:

1. **Transcript-aware input layer** – Replace or augment the mature protein sequence input with paired genomic/transcript data (available from RefSeq, Ensembl, or GTEx). Delay embeddings are built on the longer T signal before splicing.

2. **Intron-aware delay ablation** – Systematically vary delay families on intron-containing transcripts to quantify which separations (splice-junction proximity, intron length, etc.) contribute most to structure and affinity prediction.

3. **Co-translational regulariser** – Add a causal (autoregressive) component that simulates the growing chain, using delay coordinates to enforce that early residues influence later folding while the chain is still incomplete.

4. **Joint structure–affinity–splicing objective** – Extend Equation (30) to include a splicing reconstruction term:

$$\mathcal{L} = \lambda_S \mathcal{L}_{\text{structure}} + \lambda_A \mathcal{L}_{\text{affinity}} + \lambda_B \mathcal{L}_{\text{binding}} + \lambda_{\text{splice}} \mathcal{L}_{\text{splice}}. \tag{7}$$

These extensions remain fully compatible with the six-step research programme in the main paper (Section 18). Step 1 (sequence-to-structure validation) can now be performed on both spliced and unspliced signals; Step 2 (delay-scale ablation) gains an explicit biological interpretation via intron positions.

5 Open questions for future researchers

- How much additional predictive power does the full primary transcript provide over the mature protein sequence alone, especially for proteins with complex alternative splicing?
- Can learned delay families recover known biological periodicities (e.g., exon length distributions, intron-mediated pausing) without explicit supervision?
- To what extent do intron-containing delay embeddings improve generalisation to novel folds or ligand scaffolds outside the training distribution?
- Can the model be trained end-to-end to predict both splicing patterns and binding affinity from genomic sequence, thereby closing the loop from DNA to phenotype?
- What new diagnostics (e.g., delay-attention maps) become possible once the construction signal is treated as a true dynamical trace?

Part 2 does not alter the core programme of the main paper; it makes its biological grounding more precise. The Takens-based approach was already motivated by the sequential nature of construction signals. Recognising transcription, splicing, and translation as literal time series simply equips the framework with a richer observable and a deeper set of delay scales with which to reconstruct hidden geometric and functional constraints.