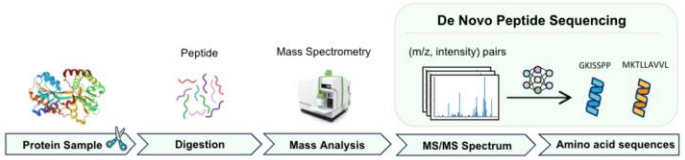
Bidirectional Representations Augmented Autoregressive Biological Sequence Generation: Application in De Novo Peptide Sequencing

Xiang Zhang, Jiaqi Wei, Zijie Qiu, Sheng Xu, Zhi Jin, ZhiQiang Gao, Nanqing Dong, Siqi Sun

Motivation

- **Bidirectional Information**: Autoregressive (AR) models, are limited in many biological by their unidirectional nature, failing to capture global bidirectional token dependencies.
- **Scalable Generation**: Non-Autoregressive (NAR) models offer holistic, bidirectional representations but face challenges with generative coherence and scalability.

De Novo Peptide Sequencing Pipeline



Core Design of CrossNovo

(1) Cross-Decoder NAT Knowledge Transfer

Algorithm 1 CROSSNOVO: AT Fine-tuning with Cross-Decoder NAT Knowledge Transfer

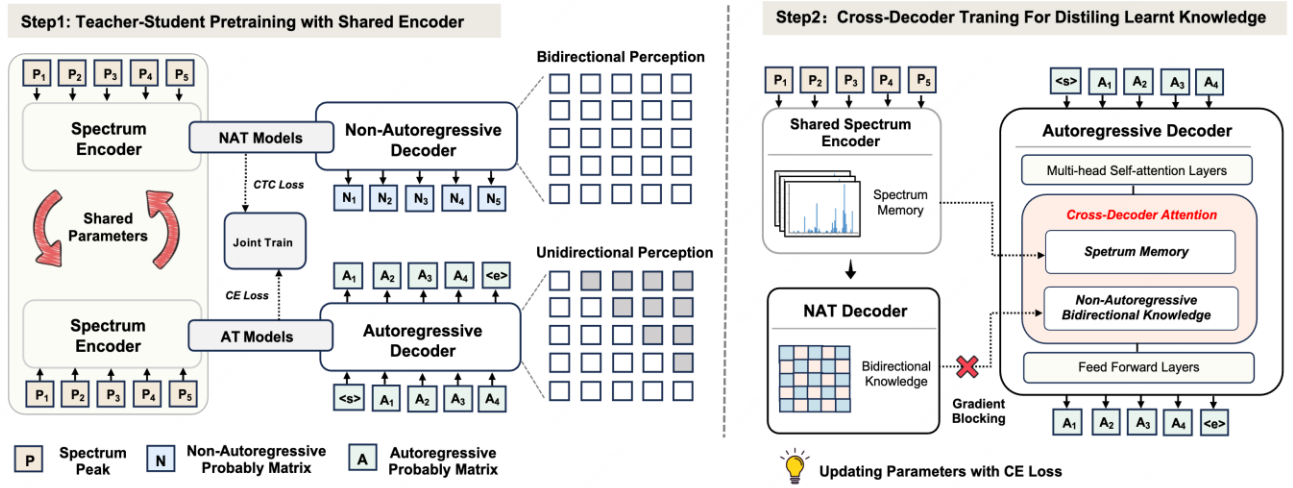
Inputs: Dataset $\mathcal{D}$; Pre-trained parameters $(\theta_{\text{enc}}, \theta_{\text{AT}}, \theta_{\text{NAT}})$ from Stage 1; Fine-tuning epochs E_{ft}; Learning rate η_{ft}; Max NAT length $T_{\max}$.
if $E_{\text{ft}} > 0$ **then**
  Freeze parameters θ_{enc} and θ_{NAT}.
  for epoch $e = 1$ **to** E_{ft} **do**
    for each $(\mathcal{S}, \mathbf{A})$ in $\mathcal{D}$ **do**
      $\mathbf{E}^{(b)} \leftarrow \text{Encoder}(\mathcal{S}; \theta_{\text{enc}})$ ▷ Use frozen encoder
      $\mathbf{V}^{(L')} \leftarrow \text{NATForward}(\text{PositionalEmbeddings}(T_{\max}), \mathbf{E}^{(b)}; \theta_{\text{NAT}})$ ▷ NAT features from frozen NAT decoder
      $\mathbf{V}^{(L')}_{\text{blocked}} \leftarrow \mathbb{GB}(\mathbf{V}^{(L')})$ ▷ Cross-decoder gradient blocking
      $\mathbf{C}_{\text{aug}} \leftarrow [\mathbf{V}^{(L')}_{\text{blocked}} \oplus \mathbf{E}^{(b)}]$ ▷ Augmented context for AT, cf. Eq. 8, 10 (with distinct positional encodings)
      $\mathcal{L}_{\texttt{AT-ft}} \leftarrow \text{ComputeATLossAugmented}(\mathbf{A}, \mathbf{C}_{\text{aug}}; \theta_{\text{AT}})$
      Update θ_{AT} using $\nabla_{\theta_{\text{AT}}} \mathcal{L}_{\texttt{AT-ft}}$ with η_{ft}.
Return Fine-tuned parameters θ_{AT} (and unchanged $\theta_{\text{enc}}, \theta_{\text{NAT}}$).

(2) Cross-Decoder Gradient Blocking for Stable Learning

- This isolation lets the AT decoder learn from the NAT branch stably, boosting model's performance.

$$\mathbf{h}_t^{\text{update}} = \text{CrossAttn}\left(\mathbf{h}_t'^{(l)}, \left[\mathbb{GB}\left(\mathbf{V}^{(L')}_{\text{p}\{1:T_{\max}\}}\right) \oplus \mathbf{E}^{(b)}_{\text{p}\{T_{\max}+1:T_{\max}+k\}}\right]\right)$$

Overview of CrossNovo Framework



- **Step 1** involves joint training with a shared encoder in a multitask learning framework, enabling the simultaneous training of Autoregressive and Non-Autoregressive decoders.
- **Step 2** introduces a novel knowledge distillation process, transferring insights from the NAT module to the AT module through a cross-decoder attention mechanism
- **Cross-decoder gradient blocking** is employed throughout to optimize the training process.

Core 9-species-v1 benchmark

Metrics	Architect	Methods	Mouse	Human	Yeast	M.mazei	Honeybee	Tomato	Rice bean	Bacillus	C. bacteria	Average
Amino Acid Precision	DB	Peaks	0.600	0.639	0.748	0.673	0.633	0.728	0.644	0.719	0.586	0.663
	NAT	Prime.	0.784	0.729	0.802	0.801	0.763	0.815	0.822	0.846	0.734	0.788
	AT	Deep.	0.623	0.610	0.750	0.694	0.630	0.731	0.679	0.742	0.602	0.673
		Point.	0.626	0.606	0.779	0.712	0.644	0.733	0.730	0.768	0.589	0.687
		Casa.	0.689	0.586	0.684	0.679	0.629	0.721	0.668	0.749	0.603	0.667
		Insta.	0.703	0.636	0.691	0.712	0.660	0.732	0.711	0.739	0.619	0.689
		Casa.V2	0.760	0.676	0.752	0.755	0.706	0.785	0.748	0.790	0.681	0.739
		Helix.	0.765	0.665	0.768	0.784	0.757	0.721	0.793	0.816	0.681	0.750
		Contra.	0.798	0.771	0.797	0.799	0.745	0.810	0.807	0.828	0.711	0.785
		Ours	**0.816**	**0.800**	**0.814**	**0.826**	**0.785**	**0.830**	**0.831**	**0.856**	**0.740**	**0.811**
Peptide Recall	DB	Peaks	0.197	0.277	0.428	0.356	0.287	0.403	0.362	0.387	0.203	0.322
	NAT	Prime.	0.567	0.574	0.697	0.650	0.603	**0.697**	0.702	0.721	**0.531**	0.638
	AT	Deep	0.286	0.293	0.462	0.422	0.330	0.454	0.436	0.449	0.253	0.376
		Point.	0.355	0.351	0.534	0.478	0.396	0.513	0.511	0.518	0.298	0.439
		Casa.	0.426	0.341	0.490	0.478	0.406	0.521	0.506	0.537	0.330	0.448
		Helix.	0.483	0.392	0.568	0.560	0.473	0.560	0.623	0.596	0.388	0.517
		Insta	0.471	0.455	0.559	0.528	0.466	**0.732**	0.564	0.576	0.416	0.530
		Casa.V2	0.483	0.446	0.599	0.557	0.493	0.618	0.589	0.622	0.446	0.539
		Contra.	0.567	0.622	0.674	0.630	0.576	0.672	0.677	0.688	0.486	0.621
		Ours	**0.596**	**0.661**	**0.698**	**0.660**	**0.610**	0.695	**0.716**	**0.726**	0.518	**0.654**

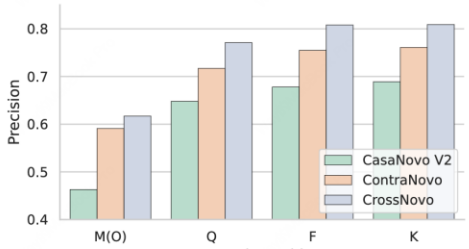
Other benchmarks

Metrics	Architect	Methods	Mouse	Human	Yeast	M.mazei	Honeybee	Tomato	Rice bean	Bacillus	C.bacteria	Average
Amino Acid Precision	NAT	Prime.	0.839	0.893	0.932	0.908	0.862	0.909	0.931	0.921	0.827	0.891
	AT	Casa.V2	0.813	0.872	0.915	0.877	0.823	0.891	0.891	0.888	0.791	0.862
		Contra.	0.839	0.920	0.919	0.896	0.848	0.898	0.913	0.901	0.807	0.882
		Ours	**0.857**	**0.937**	**0.939**	**0.920**	**0.880**	**0.914**	**0.939**	**0.927**	**0.837**	**0.906**
Peptide Recall	NAT	Prime.	0.627	0.795	0.884	0.812	0.742	**0.824**	0.837	0.849	**0.626**	0.777
	AT	Casa.V2	0.555	0.712	0.837	0.754	0.669	0.783	0.772	0.793	0.558	0.714
		Contra.	0.616	0.820	0.854	0.780	0.711	0.794	0.799	0.815	0.575	0.752
		Ours	**0.651**	**0.850**	**0.885**	**0.819**	**0.751**	0.816	**0.847**	**0.850**	0.607	**0.786**

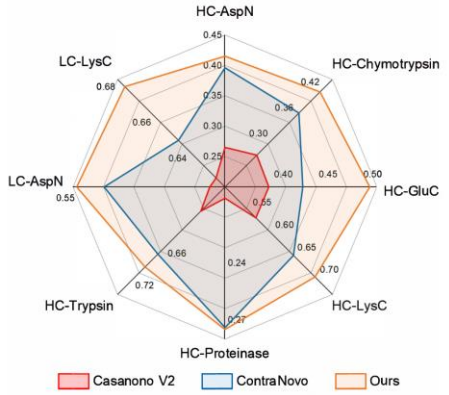
Metrics	Methods	HC AspN	HC Chymo.	HC GluC	HC LysC	HC Proteinase	HC Trypsin	LC AspN	LC LysC	Average
Amino Acid Precision	Casa.V2	0.520	0.472	0.605	0.757	0.354	0.759	0.666	0.778	0.642
	Contra.	0.580	0.565	0.642	0.790	0.348	0.787	0.702	0.793	0.676
	Ours	**0.613**	**0.617**	**0.694**	**0.814**	**0.367**	**0.803**	**0.719**	**0.807**	**0.702**
Peptide Recall	Casa.V2	0.265	0.274	0.399	0.569	0.206	0.595	0.325	0.625	0.446
	Contra.	0.396	0.372	0.437	0.653	0.274	0.675	0.499	0.646	0.529
	Ours	**0.415**	**0.421**	**0.512**	**0.701**	**0.275**	**0.699**	**0.544**	**0.676**	**0.560**

More Analysis

Performance on Amino Acids with Similar Masses



Downstream Tasks



Ablation of Core Designs

Gradient Blocking	Cross Decoder	Shared Encoder	Amino acid Precision	Peptide Precision
		✓	0.795	0.643
	✓	✓	✗	✗
✓	✓		0.698	0.546
✓	✓	✓	**0.811**	**0.654**