

JEPA-DNA: ゲノム基盤モデルにおける予測型世界モデルの構築

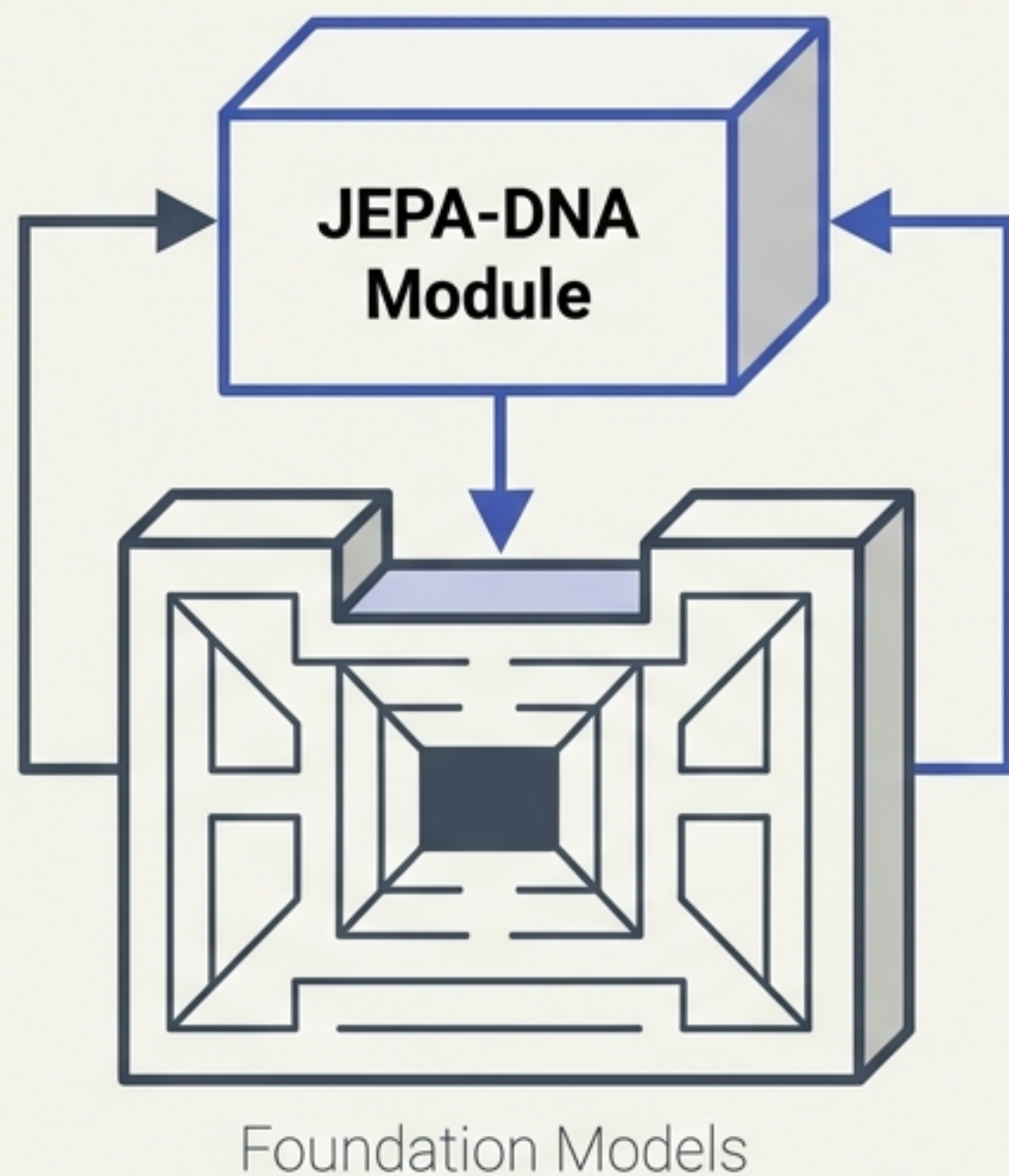
ジョイント埋め込み予測アーキテクチャによる
「粒度の罫」の克服とマルチオミクスへの展開

[FRAMEWORK]: Model-Agnostic Continual Pre-training

[CORE_TECH]: Non-contrastive Learning / VICReg

[DEVELOPER]: NVIDIA / AMI Labs

ゲノム言語モデル（gLMs）の限界を打破する継続事前学習フレームワーク



アーキテクチャの独立性（Model-Agnostic）

既存のゲノム言語モデル（DNABERT2, HyenaDNA等）のバックボーンを変更することなく、後付けで「潜在セマンティックグラウンディング」を統合。

学習パラダイムの転換

塩基レベルの「生成型（Generative）」から、グローバルな意味論を捉える「予測型（Predictive）」へのパラダイムシフト。

表現の最適化

ノイズに対する高い頑健性と、下流タスク（変異影響予測など）における極めて高い線形分離可能性（Linear-separability）を実現。

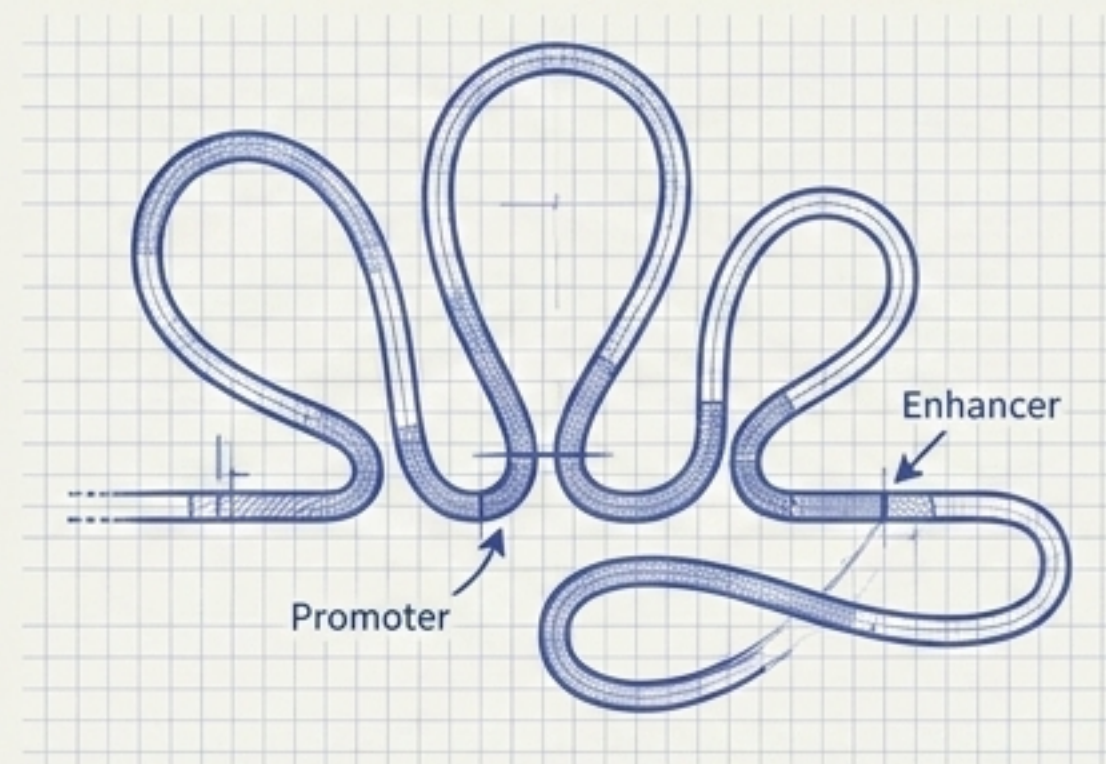
既存モデルが陥る「粒度の罠 (Granularity Trap)」



ミクロな文法 (Micro-grammar) 従来のアプローチ

マスク言語モデル (MLM) や次トークン予測 (NTP) は、コドン境界やスプライス部位など、局所的な「スペルチェック」に過剰適合する。

VS

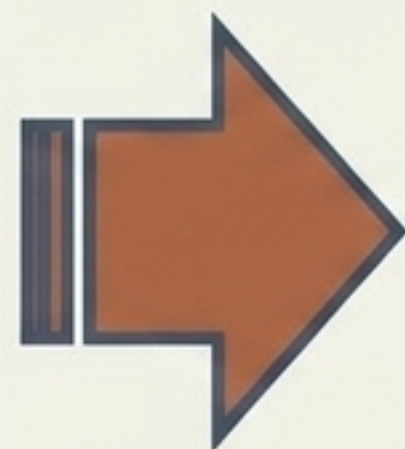
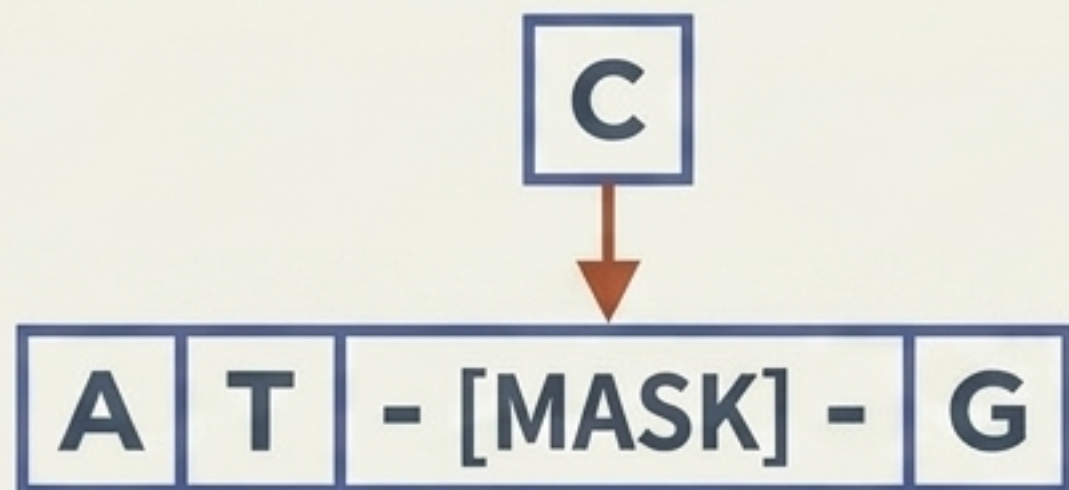


マクロな意味論 (Macro-semantics) 生物学の真実

実際のゲノム機能（プロモーター、エンハンサー、立体配置）は、数万塩基に及ぶ広域コンテキストに依存する。ミクロな復元では、このグローバルな意味論を捉えきれない。

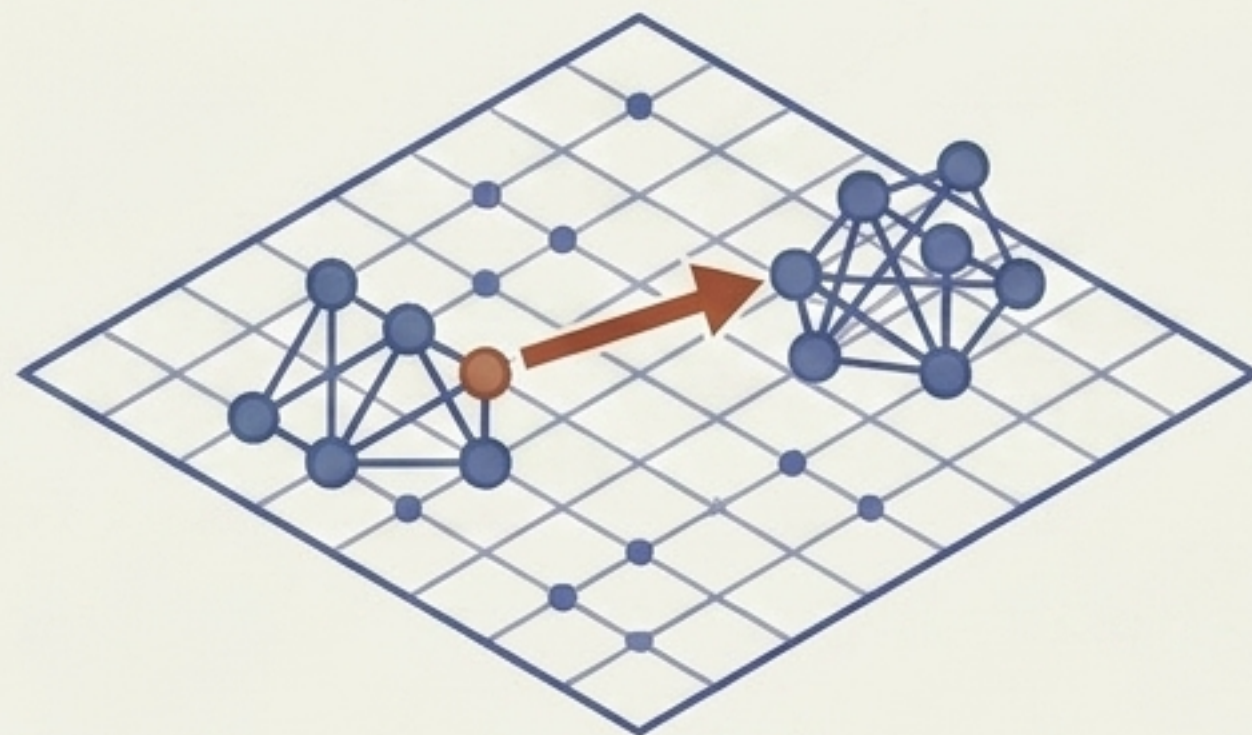
予測型世界モデルへの進化: JEPA-DNAのアプローチ

生成的再構築 (Generative Reconstruction)



ピクセルや塩基の直接的な復元。高周波ノイズ（些末な変異やエラー）に計算リソースを浪費。

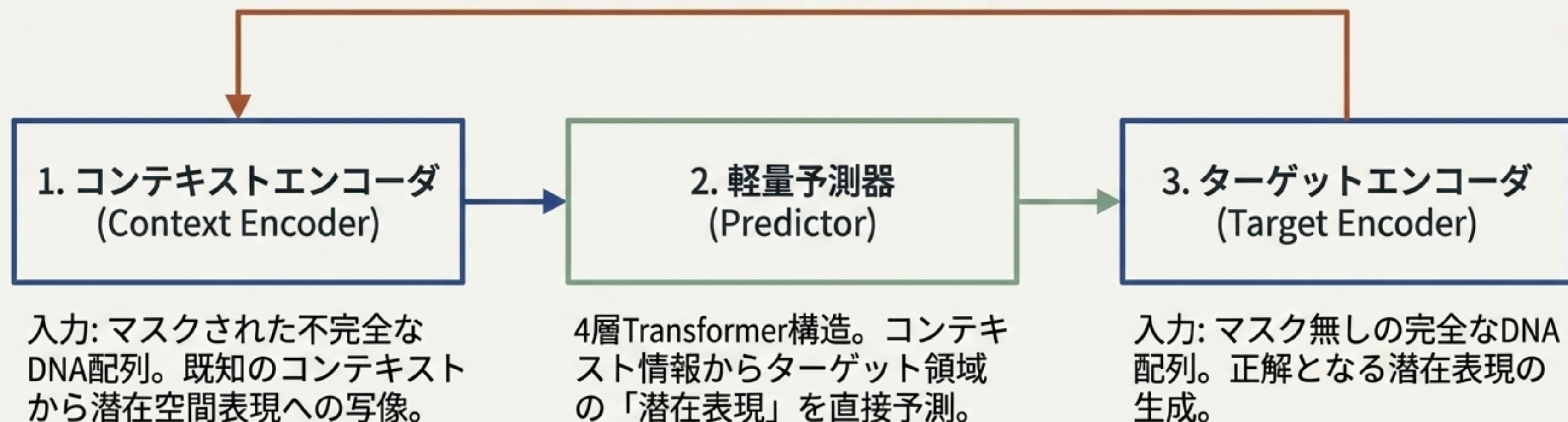
潜在空間でのセマンティック整列 (Semantic Alignment)



Yann LeCunが提唱した「予測型世界モデル」。局所的な塩基の復元を放棄し、潜在空間上でのマクロな意味の予測に学習シグナルを集中。

コアアーキテクチャ: 非対称デュアルエンコーダ構造

EMA Feedback Loop: ターゲットエンコーダの重みは、コンテキストエンコーダの重みの指数移動平均 (EMA) として同期。表現崩壊の防止を担う。



最適化エンジン: 複合損失関数と表現崩壊の回避

[Pillar 1] ネイティブ生成損失 (局所文法の維持)

$$\min_{\theta, \phi} \mathcal{L}_{total} = \lambda_1 \mathcal{L}_{llm} + \lambda_2 \mathcal{L}_{jepa} + \lambda_3 \mathcal{L}_{var} + \lambda_4 \mathcal{L}_{cov}$$

[Pillar 4] 共分散正則化
(特徴の独立性)

すべての入力が同一のベクトルに収束する「表現の崩壊 (Representation Collapse)」を、ネガティブサンプル無しで完全に回避するための数理的設計。

[Pillar 2] 潜在予測損失
(マクロ意味論の獲得)

[Pillar 3] 分散正則化
(バッチ内の多様性)

The Four Pillars of Loss: 複合損失関数の解剖学

ネイティブLLM損失

$$\mathcal{L}_{llm} \quad \mathcal{L}_{llm} = - \sum_{i \in \text{masked}} \log P(x_i | x_{\setminus i})$$

バックボーン本来の生成目的関数。局所的な塩基塩基の文法規則（ミクロ構造）を維持。

JEPA潜在予測損失

$$\mathcal{L}_{jepa} \quad \mathcal{L}_{jepa} = 1 - \frac{z \cdot z'}{\|z\| \|z'\|}$$

マスク領域の機能的意味論を直接予測。コンテキストとターゲット表現間のコサイン類似度を最大化。

分散損失 (Variance)

$$\mathcal{L}_{var} \quad \mathcal{L}_{var} = \frac{1}{d} \sum_{j=1}^d \max(0, \gamma - \sigma(z_j))$$

バッチ内の各次元の多様性を担保し、全サンプルが同一の埋め込みに収束する「崩壊」を防止。

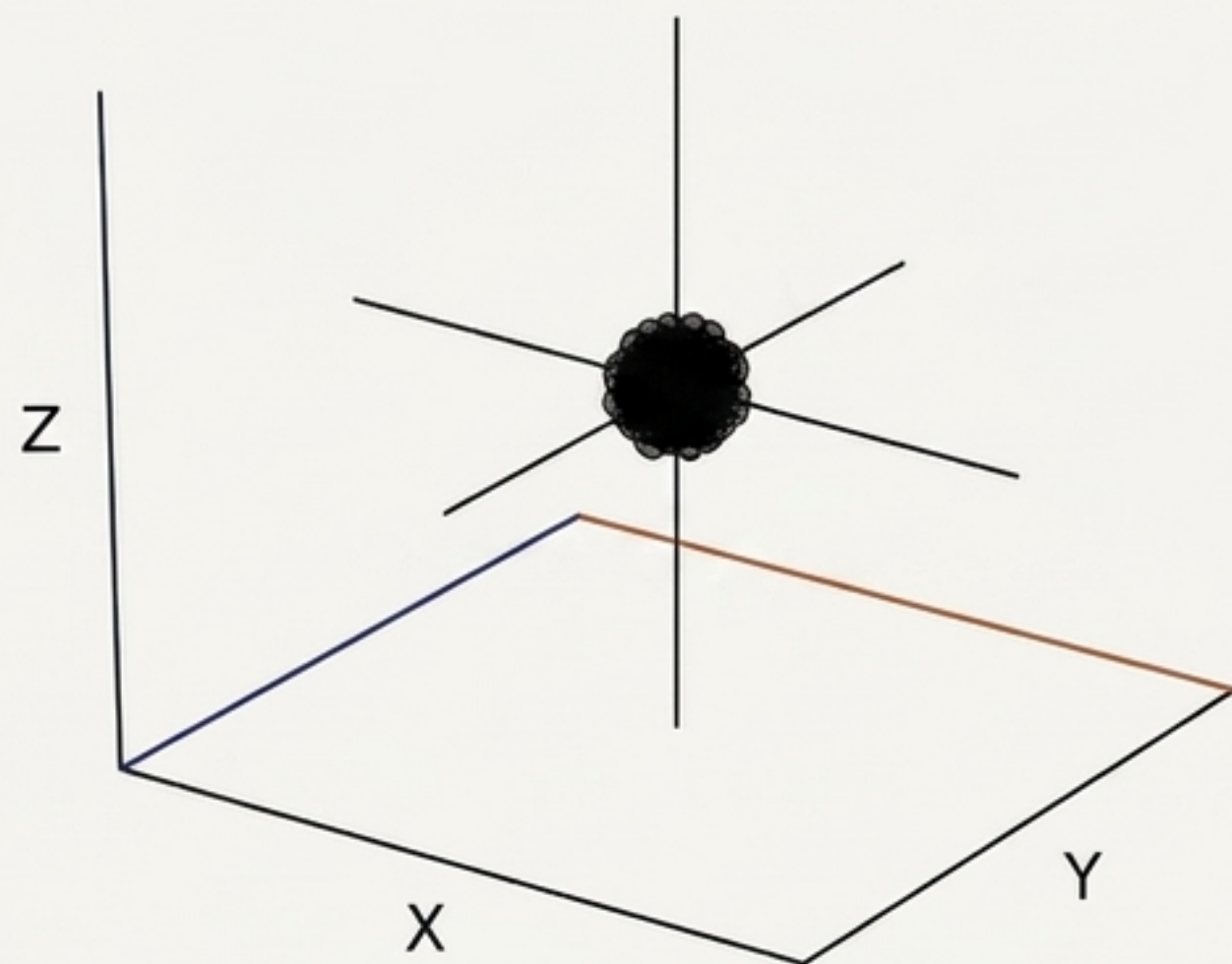
共分散損失 (Covariance)

$$\mathcal{L}_{cov} \quad \mathcal{L}_{cov} = \frac{1}{d} \sum_{i \neq j} [C]_{i,j}^2$$

次元間の相関を排除。各次元が直交する独立した生物学的特徴を表現するように強制。

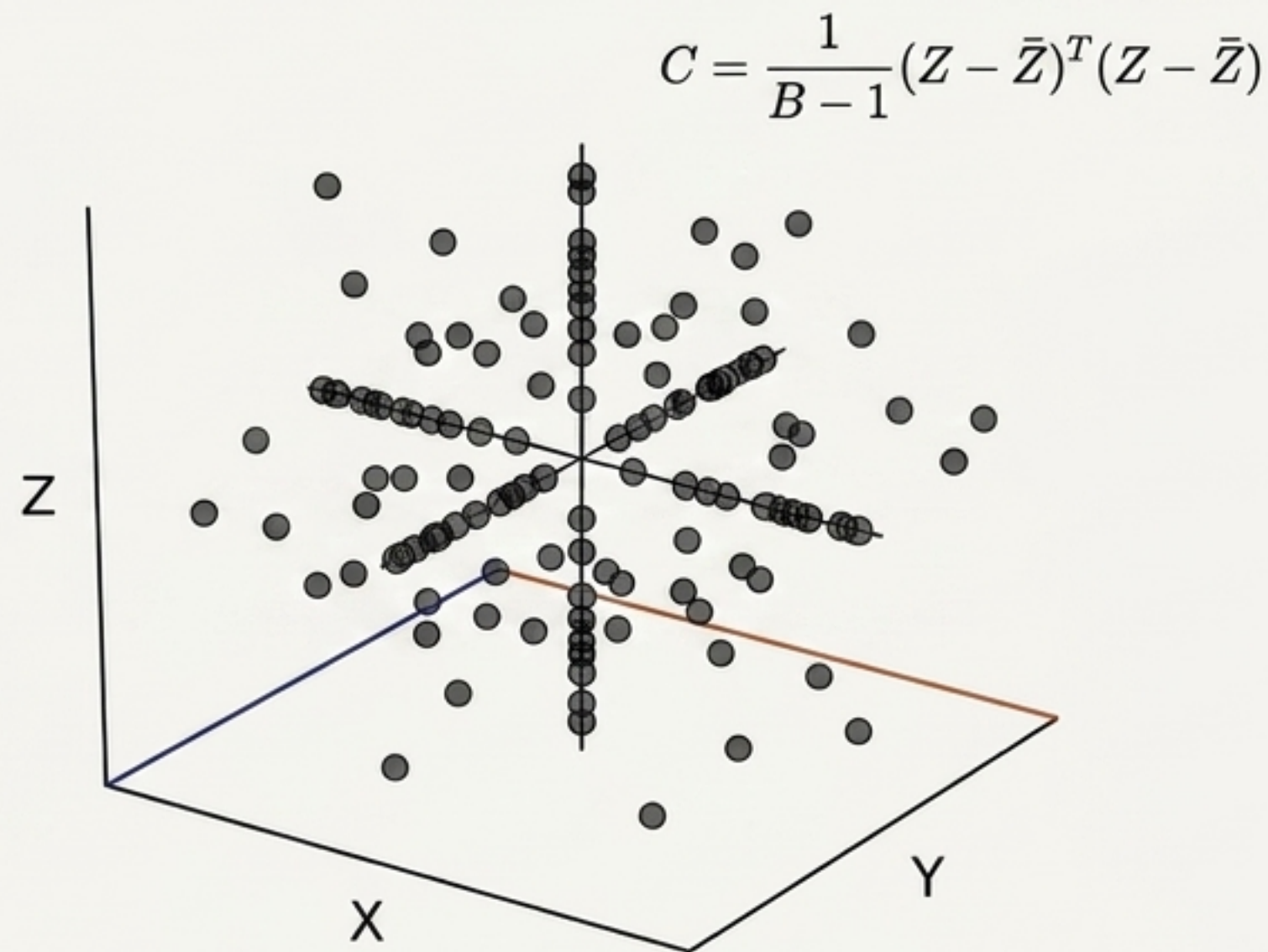
VICRegアプローチ: ネガティブサンプルに依存しない崩壊防止

表現の崩壊 (Without Regularization)



単純な非対照学習の罫。ネットワークが全ての入力に対して同一の定数を出し、情報量がゼロになる。

VICRegによる最適化空間 (With Var/Cov Loss)



$$C = \frac{1}{B-1}(Z - \bar{Z})^T(Z - \bar{Z})$$

分散 (Variance) が空間の広がり確保し、共分散 (Covariance) が特徴間の直交性を強制。意味的に豊かな表現空間を構築する。

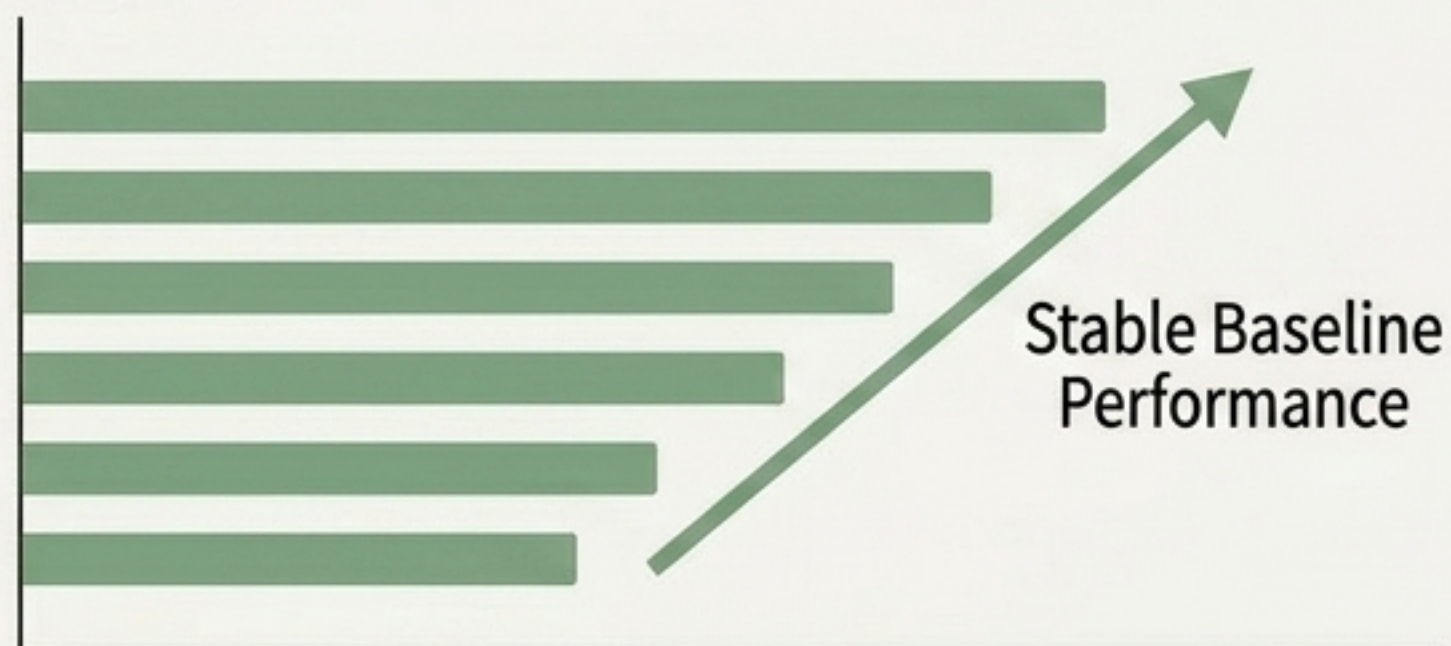
リリースモデルファミリーと適用マトリクス

比較項目	NV-JEPA-DNA-DNABERT2	NV-JEPA-DNA-HyenaDNA	NV-JEPA-DNA-NTv3
パラメータ数	117M	600K	多目的 Transformerベース
アーキテクチャ特性	高密度Attention / MLM履歴	SSM / ロング畳み込み (準2次時間)	マルチスケール K-mer
入力仕様	BPE (スパンベース)	単一塩基 (最大8,192 bp長文脈)	K-merトークン化
最適適用領域	変異影響の高精度予測、 発現調節因子の注釈	長距離ゲノム相互作用、 配列摂動スコアリング	進化保存領域の特定、 複合制御部位の機能予測

GFMBench-APIによるベンチマーク実証

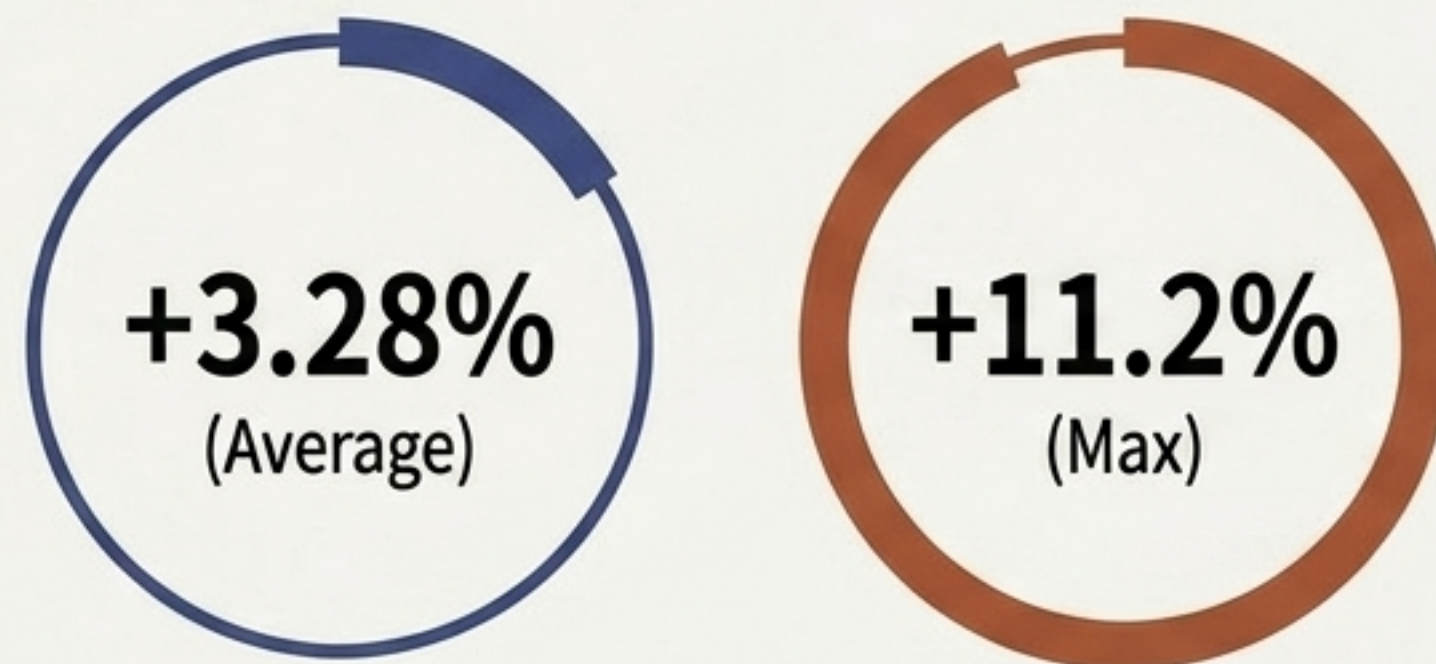
【評価概要】 17のゲノム評価タスク（プロモーター予測、転写因子結合、変異影響など）での標準評価。

Track 1: 線形プロービング (9 Tasks)



DNABERT-2バックボーンにおいて、生成タスク本来の精密さを犠牲にすることなく、制御領域のセマンティック機能識別などにおいて線形分離性能が有意に向上。

Track 2: ゼロショット評価 (8 Tasks)



HyenaDNAバックボーンにおいて、追加訓練なしで平均AUROCが約3.28%向上。特定の下流タスクでは最大+11.2%の劇的な性能改善を達成。

実装・実行環境の構築プロトコル (Deployment Pipeline)

Step 1: Environment & Caching

PyTorch 2.6.0 / Python 3.10環境。
HFキャッシュパスのエクスポート。

```
export HF_HOME="/path/to/hf_home"
```

Step 2: GFMBench-API Integration

NVIDIA標準評価ミドルウェアのクローンと
PYTHONPATHの統合。

```
export PYTHONPATH="/path/to/JEPA-DNA:/path/to/GFMBench-api"
```

Step 3: Checkpoint Retrieval

【重要】 transformers破損回避のため
CLIバージョンを固定。

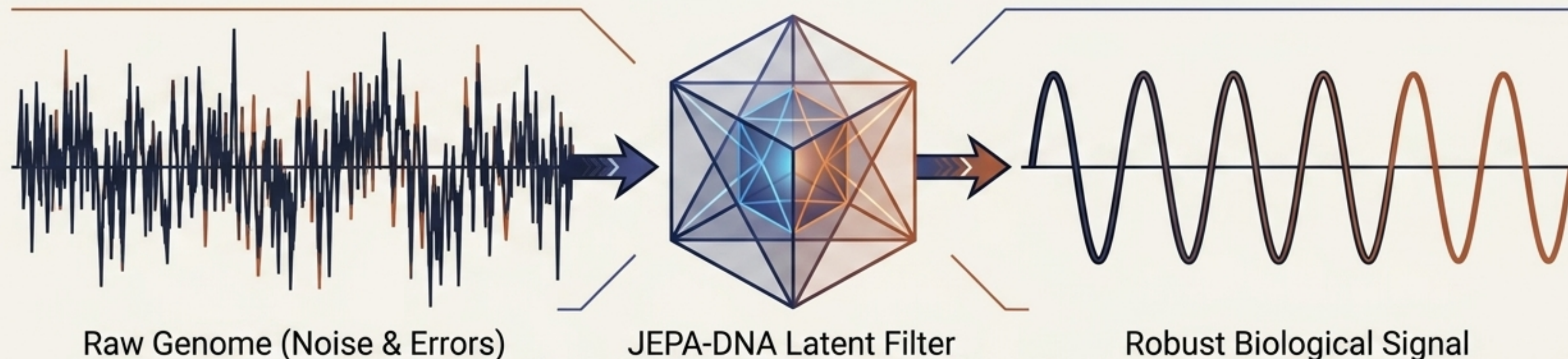
```
pip install "huggingface_hub[cli]==0.36.2"
```

Step 4: Continual Pre-training

params.jsonにバックボーンを定義し実行。

```
python ./jepa_dna/run_jepa_pretrain.py \  
/path/to/training_dir
```

数理から生物学へ: 進化的頑健性と変異影響予測 (VEP)



高周波ノイズのフィルタリング (Evolutionary Robustness)

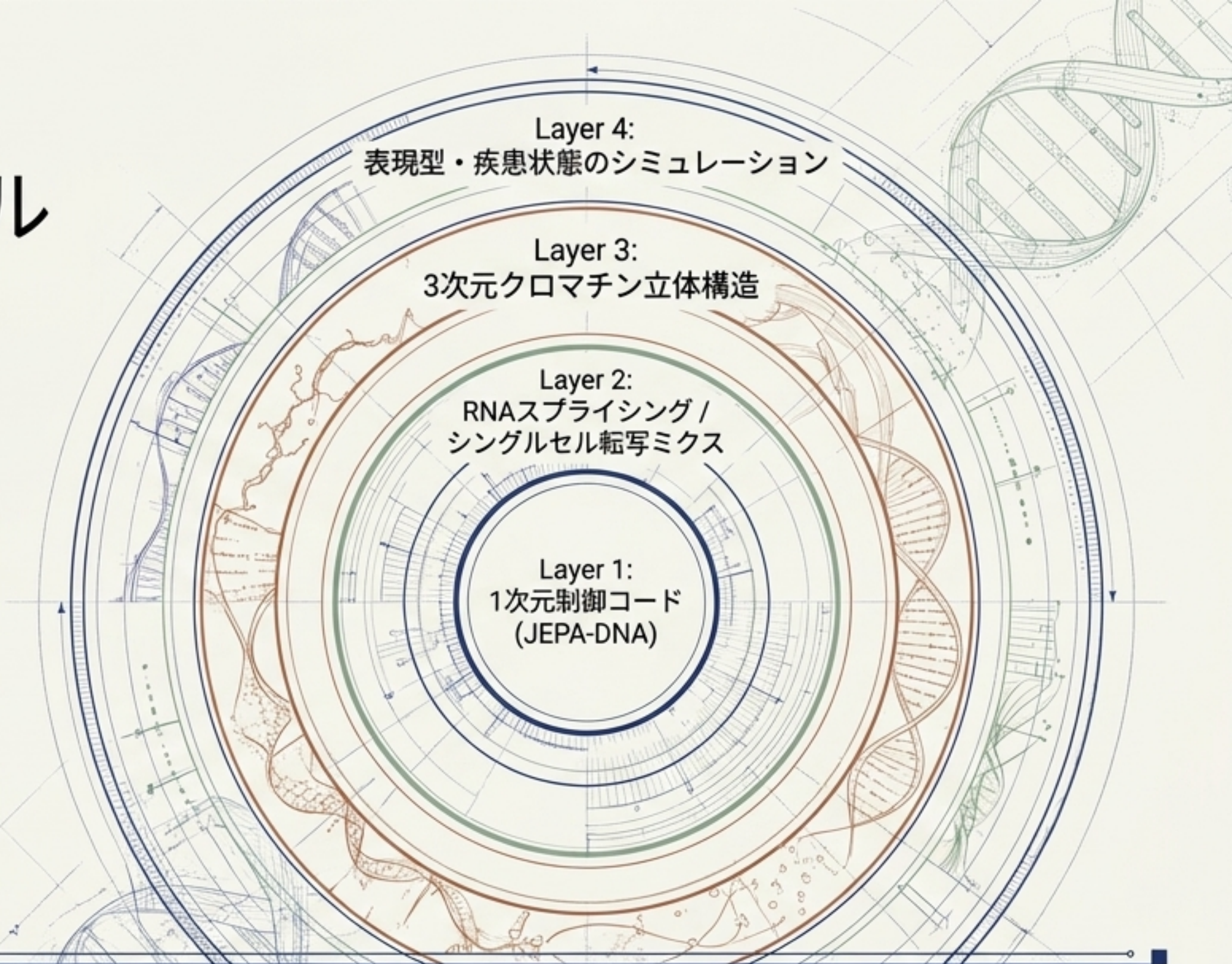
ゲノムデータの持つ自然な個体差やシーケンスエラー（配列レベルの瑣末な差分）を、JEPAの潜在空間予測が自然に「ノイズ」として無視する。異なる系統間でも偏りのない中立な解釈が可能。

変異影響予測への即応 (Variant Effect Prediction)

実験的検証が困難な数百万の非コーディング領域バリエーション（SNVやインデル）に対し、ゼロショットで配列変化に伴う潜在的な空間摂動スコアを安定出力。強力なデジタルフィルターとして機能。

ゲノムAIの次なる地平: マルチオミクス世界モデル への昇華

JEPAアーキテクチャの基本概念は、
1次元のDNAシーケンスに留まらない。



AIの役割は「分類支援ツール」から、生物学的ダイナミクス全体をシミュレーションする「予測世界モデル」へ。生体モダリティの水平展開を支える究極の基盤となる。