Stanford RNA 3D Folding 2: Sıfırdan Leaderboard Zirvesine Kapsamlı Rehber

RNA sekansından 3D yapı tahmini yapan bu Kaggle yarışmasında **$75,000** ödül havuzu için yarışıyorsunuz. NVIDIA Developer Değerlendirme metriği **TM-score** (0-1 arası, yüksek iyi) Competitions ve her sekans için **5 farklı yapı tahmini** sunmanız gerekiyor. Competitions En kritik başarı faktörü: RibonanzaNet2 foundation modelini fine-tune etmek NVIDIA Developer ve ensemble stratejileri uygulamak. Yarışma 22 Mayıs 2025'te sona eriyor givemechallenge ve sadece CASP16 kapanış tarihi (30 Eylül 2024) öncesi veriler kullanılabiliyor. Competitions

---

Yarışmanın tam olarak ne istediği

Stanford Das Lab tarafından düzenlenen bu yarışma, NVIDIA Developer RNA molekülünün **sadece sekansından** 3D uzaysal yapısını tahmin etmenizi istiyor. Competitions Tahmin edilecek değer her nükleotid için **C1' atomu koordinatları** (x, y, z) Angstrom cinsinden. Her test sekansı için **5 adet bağımsız 3D yapı tahmini** zorunlu Competitions - sistem bunların en iyisini değerlendiriyor. Competitions

TM-score formülü doğrudan RNA yapı benzerliğini ölçüyor:

$$\text{TM-score} = \max\left(\frac{1}{L_{ref}} \times \Sigma\left[\frac{1}{1+(d_i/d_0)^2}\right] \right)$$

Burada Lref referans yapıdaki rezidü sayısı, di hizalanmış rezidüler arası mesafe, d0 ise uzunluğa bağlı normalize edici faktör. **30 rezidünün altındaki kısa RNA'lar için d0 sabit değerler alıyor** (Lref<15 için d0=0.3). Competitions Bu detay kısa sekans tahminlerinde kritik.

Ödül dağılımı dikkat çekici: 1. sıra $45,000, 2. sıra $15,000, 3. sıra $10,000. NVIDIA Developer Ayrıca **Early Sharing Prize** olarak VFOLD_human_expert benchmark'ını geçen ilk 2 public notebook'a $2,500 veriliyor. givemechallenge Competitions

---

Veri formatı ve submission yapısı

Input dosyası (test_sequences.csv) iki sütun içeriyor: ID (sekans kimliği) ve sequence (A, C, G, U harflerinden oluşan RNA dizisi). Encoding basit: {'A':0, 'C':1, 'G':2, 'U':3}. Medium

Submission formatı kritik - doğru yapıda olmazsa sıfır puan:

```csv
ID,resname,resid,x_1,y_1,z_1,x_2,y_2,z_2,x_3,y_3,z_3,x_4,y_4,z_4,x_5,y_5,z_5
R1107_1,G,1,-7.561,9.392,9.361,-7.421,9.112,9.001,-7.201,9.532,9.121,-7.681,9.252,9.461,-7.301,9.023,8.932
```

Competitions

Her satır bir nükleotidi temsil ediyor. ID formatı {seq_id}_{pozisyon}, resname nükleotid harfi (G/A/C/U), resid 1'den başlayan pozisyon numarası. **15 koordinat sütunu** var: 5 tahmin × 3 koordinat (x, y, z). Competitions

Ek veri kaynakları yarışmada mevcut:

- Ribonanza_bpp_files/ — EternaFold Base Pair Probability matrisleri
- rhofold_pdbs/ — RhoFold tarafından üretilmiş PDB dosyaları (baseline olarak kullanılabilir)
- supplementary_silico_predictions/ — In-silico tahminler

Kazanan yaklaşımlar ve en başarılı teknikler

Önceki Ribonanza yarışmasının **top 6 çözümünün tamamı** advanced transformer mimarileri kullandı. FEBS Network febs İşte kanıtlanmış stratejiler:

1. sıra (Team vigg - ArmNet): Transformer + 1D Convolution hibrit mimarisi. nih Her attention bloğundan sonra 1D conv modülü eklendi. EternaFold BPP matrislerini attention bias olarak kullandılar. PubMed Central nih **15 model ensemble** ile en iyi MAE'yi elde ettiler. PubMed

2. sıra (Hoyeol Sohn): Benzer 1D conv + attention yapısı, BPP features kullanımı.

3. sıra (Team Arinka): AlphaFold2 benzeri "Twin Tower" mimarisi — sequence track'ten pair track'e bilgi akışı. İlginç şekilde **BPP kullanmayan** bir model varyantı da başarılı oldu. PubMed Central nih

4. sıra (yu4u): PyTorch Lightning ile 5-fold CV, DDP multi-GPU, mixed precision (FP16), signal-to-noise threshold filtresi, GitHub pseudo-labeling ve weighted ensemble. GitHub

---

Kaggle GPU'da çalışacak en iyi modeller

Kaggle notebook'larında **16GB VRAM** (Tesla P100 veya 2×T4) ve **haftalık 30 saat** GPU kotası var. Ultralytics GMI Cloud Bu kısıtlamalar altında en uygun modeller:

RhoFold+ (⭐ Birincil Öneri)

- **Inference süresi:** ~0.14 saniye (MSA olmadan) News-Medical PubMed
- **GPU memory:** 8-12GB — Kaggle uyumlu
- **Özellik:** Tek sekans modunda çalışabilir, MSA opsiyonel GitHub
- **Erişim:** github.com/ml4bio/RhoFold, HuggingFace'de weights mevcut
- **Neden ideal:** Nature Methods 2024'te SOTA, RNA-FM foundation model üzerine kurulu Nature News-Medical

DRfold (MSA-free alternatif)

- **Avantaj:** MSA gerektirmiyor — pipeline çok basitleşiyor
- **GPU memory:** 6-10GB
- **Dezavantaj:** L-BFGS minimizasyonu nedeniyle dakikalar sürebilir
- **Erişim:** zhanggroup.org/DRfold

trRosettaRNA2 (Verimli seçenek)

- **Training:** Tek A100'de 12 günde eğitilebilir
- **GPU memory:** 4-8GB
- **Özellik:** Res2Net-enhanced transformer, en düşük parametre/performans oranı bioRxiv

RibonanzaNet2 (Yarışmanın temel modeli)

- **Parametre:** 100M NVIDIA Developer
- **Özellik:** 40M RNA sekansı üzerinde pre-trained nvidia NVIDIA Developer
- **Fine-tuning:** Das Lab tarafından özellikle bu yarışma için tasarlandı NVIDIA Developer
- **Erişim:** kaggle.com/models/shujun717/ribonanzanet2/PyTorch/alpha

AlphaFold3'ü KULLANMAYIN — 40-80GB GPU memory gerektirir ve RNA için optimize değil.

Tartışmalardan çıkan altın değerinde ipuçları

BPP Features kritik: EternaFold'dan elde edilen Base Pair Probability matrisleri, attention mekanizmasına bias olarak eklendiğinde dramatik performans artışı sağlıyor. Top Kaggle çözümlerinin hepsi bunu kullandı. FEBS Network +2

Sequence Flip augmentation: RNA sekanslarını hem 5'→3' hem 3'→5' yönünde okuyarak veri çeşitliliği artırılıyor. RibonanzaNet ve RNAdegformer'da kanıtlanmış iyileştirme. nih

Two-stage training: İlk aşamada SN>0.5 ile tüm veri, ikinci aşamada SN>1.0 ile yüksek kaliteli veri ve düşük learning rate (2e-4). Cosine annealing schedule zorunlu. GitHub

Pseudo-labeling güçlü: Top 3 Kaggle tahminlerinden türetilen pseudo label'lar ile pre-training, ardından gerçek etiketlerle fine-tuning. RibonanzaNet'in doğruluğunu kanıtlanmış şekilde artırdı. PubMed Central Danling

Length generalization önemli: Eğitim verisi 115-206 nt, test verisi 207-457 nt. Daha uzun sekanslar için generalize edemeyen modeller private leaderboard'da çökecek. nih

---

Yeni başlayanlar için adım adım strateji

Hafta 1-2: Foundation

1. **RibonanzaNet2'yi indirin** — Kaggle Models'dan alpha release
2. **Baseline notebook çalıştırın** — Stanford RNA 3D Folding EDA + Baseline notebook'unu fork edin
3. **Veri yapısını anlayın** — test_sequences.csv, submission format, TM-score hesaplama
4. **İlk submission yapın** — RhoFold PDB'lerinden baseline tahminler

Hafta 3-4: Model Geliştirme

5. **BPP features hesaplayın** — EternaFold veya LinearPartition kullanın
6. **Transformer + Conv mimarisi kurun** — Attention bloklarından sonra 1D conv
7. **Mixed precision training** — `trainer.precision=16` zorunlu
8. **5-fold CV başlatın** — Her fold için ayrı model

Hafta 5-6: İleri Teknikler

9. **Pseudo-labeling uygulayın** — Top submission tahminlerinden
10. **Ensemble oluşturun** — Minimum 5 model, weighted average
11. **Sequence flip augmentation** — 3'→5' ve 5'→3' yönleri

Son Hafta: Optimizasyon

12. **OOF predictions** ile ensemble ağırlıklarını optimize edin
13. **Post-processing** — Confidence-based filtering
14. **Final submission** — En iyi 5 farklı yapı tahmini

Pratik kod örnekleri

Dataset sınıfı (PyTorch):

```python
```

```python
class RNA_Dataset(Dataset):
    def __init__(self, df, Lmax=512):
        self.seq_map = {'A': 0, 'C': 1, 'G': 2, 'U': 3}
        self.Lmax = Lmax
        self.df = df

    def __getitem__(self, idx):
        seq = self.df.iloc[idx]['sequence']
        L = len(seq)

        seq_encoded = torch.zeros(self.Lmax, dtype=torch.long)
        seq_encoded[:L] = torch.tensor([self.seq_map[s] for s in seq])

        mask = torch.zeros(self.Lmax, dtype=torch.bool)
        mask[:L] = True

        return {'sequence': seq_encoded, 'mask': mask, 'length': L}
```

Submission oluşturma:

```python
def create_submission(predictions, test_df):
    """
    predictions: dict of {seq_id: np.array shape (L, 5, 3)}
    """
    rows = []
    for idx, row in test_df.iterrows():
        seq_id, sequence = row['ID'], row['sequence']
        pred = predictions[seq_id]

        for pos, nuc in enumerate(sequence):
            row_data = {'ID': f"{seq_id}_{pos+1}", 'resname': nuc, 'resid': pos + 1}
            for m in range(5):
                row_data[f'x_{m+1}'] = pred[pos, m, 0]
                row_data[f'y_{m+1}'] = pred[pos, m, 1]
                row_data[f'z_{m+1}'] = pred[pos, m, 2]
            rows.append(row_data)

    return pd.DataFrame(rows)
```

Memory-efficient training config:

```python
```

```
trainer = pl.Trainer(
    precision=16,  # Mixed precision
    strategy='ddp',  # Multi-GPU
    gradient_clip_val=1.0,
    accumulate_grad_batches=4,  # Effective batch size artırma
    max_epochs=32,
    callbacks=[EarlyStopping(monitor='val_loss', patience=5)]
)
```

Ensemble ve post-processing teknikleri

Model ensemble stratejisi: Minimum 5 farklı model kullanın — farklı fold'lar, farklı augmentation'lar, farklı hyperparameter'lar. Weighted average için OOF (out-of-fold) skorlarını kullanın.

5 yapı tahmini için çeşitlilik: Her sekans için 5 farklı yapı gerekiyor. Stratejiler:

- Farklı random seed'lerle aynı model
- Farklı model mimarilerinden tahminler
- Temperature sampling ile çeşitlilik
- Top-K diverse structure selection

Confidence-based filtering: pLDDT (predicted LDDT) skorlarını kullanarak düşük güvenilirlikli tahminleri filtreleyin. Yüksek confidence tahminleri ensemble'a daha yüksek ağırlıkla dahil edin.

Template-based hibrit yaklaşım: İlginç bir bulgu — önceki RNA yarışmasında en iyi strateji **template-based modeling** (deep learning değil) oldu. PDB'den benzer yapıları bulup template olarak kullanmak güçlü bir strateji.

Kritik uyarılar ve yaygın hatalar

CASP16 veri sınırlaması: Sadece 30 Eylül 2024 öncesi halka açık veriler kullanılabilir. (Competitions) Bu tarihten sonraki PDB yapılarını kullanmak diskalifiye nedeni. (Competitions)

Overfitting riski: Bu yarışma "blind test" formatında — private leaderboard verileri yarışma sırasında toplanıyor. Public leaderboard'a overfit olmak tehlikeli.

Uzun sekans generalization: Test verileri eğitimden daha uzun sekanslar içeriyor. Sabit pozisyonel encoding kullanmayın, relative positional encoding tercih edin.

Submission format hatası: ID formatı, koordinat sıralaması ve sütun isimleri tam olarak belirtildiği gibi olmalı. Küçük bir format hatası sıfır puan demek.

Önerilen tam pipeline

```
1. Data Loading
   └── test_sequences.csv → RNA Dataset

2. Feature Engineering
   ├── Sequence encoding (A/C/G/U → 0/1/2/3)
   ├── BPP matrix (EternaFold)
   └── Secondary structure (RNAfold/LinearPartition)

3. Model Training (5-fold CV)
   ├── RibonanzaNet2 fine-tune
   ├── Transformer + 1D Conv
   ├── Pseudo-labeling (2nd stage)
   └── Mixed precision + gradient accumulation

4. Inference
   ├── RhoFold+ (single-seq mode) × 5 seeds
   ├── Fine-tuned RibonanzaNet2 × 5 folds
   └── DRfold baseline

5. Ensemble
   ├── Weighted average (OOF-based weights)
   └── Top-5 diverse structure selection

6. Submission
   └── submission.csv (ID, resname, resid, x_1...z_5)
```

Sonuç

Bu yarışmada başarı için **RibonanzaNet2 fine-tuning**, **BPP features**, **5+ model ensemble** ve **pseudo-labeling** kombinasyonu en kanıtlanmış strateji. Kaggle 16GB GPU kısıtlamalarında (Ultralytics) RhoFold+ (single-seq mode) ve DRfold (MSA-free) en pratik seçenekler. Early Sharing Prize için VFOLD_human_expert benchmark'ını geçen public notebook yayınlamak $2,500 kazanma şansı veriyor. (givemechallenge) (Competitions)

Kritik başarı faktörleri: (1) Uzun sekans generalization, (2) 5 çeşitli yapı tahmini stratejisi, (3) CASP16 veri sınırlamasına uyum. Template-based + deep learning hibrit yaklaşımları da keşfetmeye değer.