report

December 8, 2025

1 Finding True Earth Twins: A Human-Centered Approach to Exoplanet Habitability

1.1 Abstract

This study investigates the completeness of the Earth Similarity Index (ESI) as a metric for identifying potentially habitable exoplanets. Using a dataset of 6,060 confirmed exoplanets from NASA's Exoplanet Archive, we demonstrate that non-ESI features (orbital characteristics, stellar properties) can predict 51% of ESI variance, indicating that the traditional 5-parameter ESI is incomplete. We develop an Extended ESI incorporating 10 additional data-driven features identified through Random Forest and SHAP analysis, achieving 98.6% R^2 when predicting ESI. Our Extended ESI reveals significant ranking shifts: only 10% overlap between top 20 planets by Original vs Extended ESI, with popular "Earth-like" candidates like TRAPPIST-1 e and Proxima Cen b dropping substantially in Extended ESI rankings. These findings suggest that habitability assessment requires consideration of orbital dynamics and stellar context beyond physical planetary parameters alone.

1.2 1. Introduction

1.2.1 1.1 Motivation

Ever since childhood, I've been fascinated by the night sky: the vastness, the mystery, and the question of whether Earth is truly unique. I remember spending evenings watching *Through the Wormhole* with Morgan Freeman, *Cosmos: A Spacetime Odyssey*, and NASA documentaries on the Discovery Channel. Around 2010-2012, I even started maintaining an Excel sheet of "exoplanets," noting down planets like Kepler-22b or the "Diamond Planet" and imagining which one I'd choose if someday I had to leave Earth.

Thirteen years later, I finally have the opportunity and analytical tools to revisit this childhood curiosity through a data-driven lens. With access to large-scale astronomical data and modern data science methods, I want to systematically explore how Earth's physical and environmental conditions compare to those of other planets. The central motivation is to identify which planets come closest to being "Earth-like" through quantifiable scientific features such as temperature, gravity, atmospheric pressure, and orbital distance.

Ultimately, I hope to learn which physical parameters matter most when defining "Earth-likeliness" and whether the planets often described as "Earth twins" in popular media genuinely deserve that label when analyzed quantitatively.

Scientific Context: The search for habitable worlds beyond our solar system is one of the most compelling questions in modern astronomy. With over 6,000 confirmed exoplanets discovered to date, identifying which planets might support life requires robust, scientifically-grounded metrics. The Earth Similarity Index (ESI), developed by Schulze-Makuch et al. (2011), has become a widely-used tool for ranking exoplanets by their similarity to Earth across five key physical parameters: radius, mass, density, escape velocity, and equilibrium temperature.

However, habitability is a complex, multi-faceted phenomenon that depends on far more than these five parameters. Orbital characteristics (eccentricity, semi-major axis), stellar properties (temperature, metallicity, radius), and derived metrics (surface gravity, habitable zone membership) all play critical roles in determining whether a planet can maintain liquid water and stable climate conditions over geological timescales.

This study addresses a fundamental question: **Is the traditional ESI metric incomplete?** If other features beyond ESI's five parameters can predict Earth-like conditions, then ESI may be missing important habitability signals. We test this hypothesis by training machine learning models to predict ESI using only non-ESI features, and then develop an Extended ESI that incorporates additional data-driven features to provide a more comprehensive assessment of planetary habitability.

1.2.2 1.2 Research Objectives

This project aims to:

1. **Quantify ESI incompleteness**: Measure how well non-ESI features can predict ESI, providing evidence for or against ESI's completeness
2. **Identify missing habitability signals**: Use SHAP (SHapley Additive exPlanations) to identify which orbital and stellar features matter most for Earth similarity
3. **Develop Extended ESI**: Create an improved habitability metric that incorporates both traditional ESI parameters and data-driven features
4. **Compare rankings**: Analyze how Extended ESI changes planet rankings compared to Original ESI, identifying new candidates and re-evaluating popular "Earth-like" planets
5. **Contextualize Earth**: Understand Earth's position within the broader exoplanet population across multiple dimensions

1.2.3 1.2 Dataset and License

Data Source: NASA Exoplanet Archive (https://exoplanetarchive.ipac.caltech.edu/)

Primary Dataset: Planetary Systems Composite Parameters (`pscomppars` table)

- **Total planets**: 6,060 confirmed exoplanets
- **Data access**: Public domain via NASA's Table Access Protocol (TAP) API
- **License**: NASA data is in the public domain (U.S. Government works, Title 17, Section 105)
- **API Documentation**: https://exoplanetarchive.ipac.caltech.edu/docs/TAP/usingTAP.html

Solar System Data: NASA Planetary Fact Sheet

- **Source**: https://nssdc.gsfc.nasa.gov/planetary/factsheet/
- **License**: Public domain (U.S. Government works)

Data Completeness: The dataset contains significant missing values across many parameters. We apply a two-stage imputation strategy (deterministic mass-radius relations followed by MICE iterative imputation) to handle missing data while tracking imputation flags for uncertainty propagation.

1.3 2. Background and Related Work

1.3.1 2.1 The Earth Similarity Index (ESI)

The Earth Similarity Index, introduced by Schulze-Makuch et al. (2011), provides a quantitative measure of how similar an exoplanet is to Earth. ESI is computed as the geometric mean of normalized similarity components across five key parameters:

1. **Planetary radius** (`pl_rade`): Normalized to Earth's radius (1.0 Earth radii)
2. **Planetary mass** (`pl_masse`): Normalized to Earth's mass (1.0 Earth masses)
3. **Bulk density** (`pl_dens_calc`): Normalized to Earth's density (5.51 g/cm³)
4. **Escape velocity** (`pl_escvel_km_s`): Normalized to Earth's escape velocity (11.186 km/s)
5. **Equilibrium temperature** (`pl_eqt`): Normalized to Earth's equilibrium temperature (255 K)

Each component is calculated as: $1 - |x - x_{earth}|/|x + x_{earth}|$, where x is the planet's value and x_{earth} is Earth's reference value. The final ESI is the geometric mean of all components, producing a value between 0 and 1, where 1.0 represents an exact Earth twin.

ESI has been widely adopted in exoplanet research and public communication, but it focuses exclusively on intrinsic planetary properties. It does not consider orbital dynamics, stellar characteristics, or other factors that influence habitability.

1.3.2 2.2 Habitability Beyond Physical Parameters

Orbital Characteristics: Kasting et al. (1993) established the concept of the habitable zone (HZ) as the region around a star where a planet can maintain liquid water on its surface. The HZ depends on stellar luminosity, with inner and outer boundaries determined by the runaway greenhouse effect and complete freezing, respectively. Orbital eccentricity also matters: Rodríguez-Mozos & Moya (2025) showed that high eccentricity can destabilize climate through extreme temperature variations.

Stellar Properties: Stellar temperature, metallicity, and radius influence planetary formation, atmospheric retention, and long-term habitability. M-dwarf stars, for example, present unique challenges including tidal locking, stellar flares, and atmospheric escape.

Surface Gravity and Escape Velocity: Schulze-Makuch et al. (2011) emphasized surface gravity and escape velocity as critical for atmospheric retention. Planets with insufficient gravity cannot retain atmospheres over geological timescales, making life as we know it impossible.

Mass-Radius Relationships: Seager et al. (2007) and Chen & Kipping (2017) established empirical mass-radius relationships that constrain planetary composition. Earth-like density (~5.5 g/cm³) suggests rocky composition, which is necessary (though not sufficient) for habitability.

1.3.3 2.3 Machine Learning in Exoplanet Science

Recent work has applied machine learning to exoplanet characterization, including atmospheric retrieval, habitability prediction, and feature importance analysis. SHAP (SHapley Additive ex-

Planations) values, developed by Lundberg & Lee (2017), provide a principled way to interpret machine learning models by quantifying each feature's contribution to predictions.

1.3.4 2.4 Research Gap

While ESI is a valuable tool, no prior work has systematically tested whether ESI captures all relevant habitability signals. If non-ESI features can predict ESI with high accuracy, this would indicate that ESI is incomplete and that an extended metric incorporating additional features could improve habitability assessment.

1.4 3. Research Questions and Hypotheses

1.4.1 3.1 Research Questions

The overarching goal of this project is to understand how Earth's physical and orbital conditions compare to those of other known planets, and to identify which planet(s) come "closest" to being Earth-like in a quantitative sense.

RQ1: How does Earth compare to other planets across key physical and orbital characteristics?

For example, where does Earth sit in the distribution of planet radius, mass, surface gravity, equilibrium temperature, orbital period, and stellar insolation when considered alongside exoplanets? We examine Earth's percentile position across multiple dimensions to understand where Earth falls within the distribution of known exoplanets. This contextualizes Earth's "typicality" or "uniqueness" in the exoplanet population.

RQ2: Which planets are most similar to Earth when these characteristics are combined into a single similarity measure?

Using a subset of physically meaningful features (e.g., radius, gravity, incident flux, temperature), can we compute an "Earth similarity score" and identify top-k planets that most closely resemble Earth in this multidimensional space? We develop an Extended ESI metric that incorporates both traditional ESI parameters and data-driven features identified through machine learning. We compare these rankings to Original ESI rankings to identify new candidates and re-evaluate popular "Earth-like" planets.

1.4.2 3.2 Hypotheses

H1: Planets that are often described as "Earth-like" in popular science communication will not always rank as the closest neighbors to Earth once a multi-parameter similarity metric is applied.

Many planets receive media attention as "Earth-like" based on Original ESI or single characteristics (e.g., being in the habitable zone). However, if Extended ESI captures additional habitability signals, these popular planets may not rank as highly when considering orbital dynamics and stellar context. We test this by examining the Extended ESI rankings of well-known candidates like TRAPPIST-1 e, Proxima Cen b, and Kepler-452 b.

1.5 4. Methodology

1.5.1 4.1 Data Preprocessing Pipeline

Our analysis follows a structured pipeline implemented across multiple notebooks:

1. **Data Ingestion** (`notebooks/data_ingestion.ipynb`): Downloads exoplanet data from NASA Exoplanet Archive via TAP API and integrates Solar System planets from NASA Planetary Fact Sheet for comparative analysis.

2. **Feature Engineering** (`src/build_features.py`): Computes derived features:

   - Surface gravity and escape velocity (Schulze-Makuch et al., 2011)
   - Bulk density (Seager et al., 2007)
   - Habitable zone bounds (Kasting et al., 1993)
   - Eccentricity habitability score (Rodríguez-Mozos & Moya, 2025)

3. **Imputation** (`src/impute_dataset.py`): Two-stage imputation strategy:

   - **Deterministic reconstruction**: Mass-radius relations (Chen & Kipping, 2017)
   - **Iterative imputation**: MICE with predictive mean matching (van Buuren, 2018)
   - Earth and Solar System planets excluded from model fitting to prevent bias

4. **ESI Calculation** (`src/esi.py`): Computes Original ESI (5 parameters) and Extended ESI (5 + N additional features) using geometric mean formula.

1.5.2 4.2 Feature Importance Analysis

Why Random Forest?

- Handles non-linear relationships between features
- Provides feature importance scores
- Robust to outliers and missing values (after imputation)
- Works well with mixed feature types (continuous, binary)

Why SHAP?

- Model-agnostic interpretability
- Quantification of each feature's contribution to predictions
- Ability to identify which features matter most for Earth similarity

Method:

1. Train Random Forest to predict ESI using **only non-ESI features** (orbital characteristics, stellar properties, derived metrics)
2. Compute SHAP values to identify top predictive features
3. Remove highly correlated features (threshold: 0.90) to avoid redundancy
4. Select top N features for Extended ESI calculation

1.5.3 4.3 Extended ESI Development

Extended ESI Formula: Same geometric mean structure as Original ESI, but incorporating additional features:

$$\text{Extended ESI} = \left(\prod_{i=1}^{n} \left(1 - \frac{|x_i - x_{\text{earth}}|}{|x_i + x_{\text{earth}}|} \right) \right)^{1/n}$$

Where $n = 5$ (standard ESI features) + N (top non-ESI features identified through SHAP analysis).

Validation: Train Random Forest on ESI features + top non-ESI features to predict ESI. High R^2 (>0.9) validates that Extended ESI captures the same information as Original ESI while incorporating additional habitability signals.

1.5.4 4.4 Clustering Analysis

K-means Clustering: Applied to standardized Extended ESI features to identify groups of similar planets. Optimal k determined by silhouette score.

DBSCAN: Density-based clustering for outlier detection and noise identification.

Distance to Earth: Euclidean distance in Extended ESI feature space to identify planets most similar to Earth across all dimensions.

1.5.5 4.5 Why These Methods Are Appropriate

- **Geometric mean for ESI**: Preserves the multiplicative structure of similarity (all parameters must be Earth-like for high ESI)
- **Random Forest + SHAP**: Provides interpretable feature importance while handling complex non-linear relationships
- **MICE imputation**: Preserves relationships between variables, critical for maintaining astrophysical correlations
- **Clustering**: Validates Extended ESI by showing that Earth's cluster contains planets with high Extended ESI scores

1.6 5. Findings

The following sections present our key findings. All analysis code and detailed results are available in `notebooks/eda.ipynb`.

[29]:

```python
# Import necessary libraries
import sys
from pathlib import Path
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

# Set up paths
if Path.cwd().name == "notebooks":
    project_root = Path.cwd().parent
    sys.path.insert(0, str(project_root))
else:
    project_root = Path.cwd()
    sys.path.insert(0, str(project_root))

# Data directories
data_raw = project_root / "data" / "raw"
data_processed = project_root / "data" / "processed"
reports_dir = project_root / "reports"

# Import project modules
from src.build_features import add_derived_columns, DERIVED_COLUMNS
from src.impute_dataset import impute
from src.feature_selection import compute_feature_importance,␣
  ↪compute_extended_esi as compute_extended_esi_rf, ESI_FEATURES
from src.esi import compute_esi, compute_esi_radius_mass_only,␣
  ↪compute_extended_esi

# Set visualization style
sns.set_style("whitegrid")
plt.rcParams["figure.figsize"] = (12, 8)
```

[2]:

```
# Load processed and imputed data
raw_path = data_raw / "exoplanets_pscomppars.csv"
processed_path = data_processed / "processed_exoplanets.csv"
imputed_path = data_processed / "processed_exoplanets_imputed.csv"

# Load imputed dataset (preprocessing already completed)
df_imputed = pd.read_csv(imputed_path)
print(f"Loaded imputed dataset: {df_imputed.shape}")

# Calculate ESI on imputed data
df_imputed["esi"] = compute_esi(df_imputed)
df_imputed["esi_radius_mass"] = compute_esi_radius_mass_only(df_imputed)

print(f"ESI statistics: mean={df_imputed['esi'].mean():.4f},␣
  ↪median={df_imputed['esi'].median():.4f}")
print(f"Total planets: {len(df_imputed):,}")

# Verify Earth is in dataset
earth_mask = df_imputed["pl_name"].str.lower() == "earth"
if not earth_mask.any():
    raise ValueError("Earth not found in dataset!")
print("Earth found in dataset")
```

```
Loaded imputed dataset: (6060, 704)
ESI statistics: mean=0.3032, median=0.3082
```

```
Total planets: 6,060
Earth found in dataset
```

1.6.1 5.2 Testing ESI Incompleteness

We test whether non-ESI features can predict ESI by training a Random Forest model using only non-ESI features (orbital characteristics, stellar properties, derived metrics) to predict ESI. If $R^2 > 0.5$, this provides evidence that ESI is incomplete.

[30]:

```
results = compute_feature_importance(
    df_imputed.copy(),
    output_dir=reports_dir,
    correlation_threshold=0.90,
    min_features_required=0.5,
)

importance_df = results.importance_df
feature_matrix = results.feature_matrix
X = results.X
model = results.model
shap_values = results.shap_values
df_filtered = results.df_filtered

# Use filtered dataframe for all subsequent analysis
df_imputed = df_filtered.copy()

# IMPORTANT: Update earth_mask for filtered dataset
earth_mask = df_imputed["pl_name"].str.lower() == "earth"

print(f"\nFiltered dataset: {len(df_imputed):,} planets")

# Get non-ESI features in the order they appear in feature_matrix.columns
non_esi_features_in_order = [col for col in feature_matrix.columns if col not␣
  ↪in ESI_FEATURES and col != 'esi']

# Select features in correct order
X_non_esi = feature_matrix[non_esi_features_in_order]
y_true = feature_matrix['esi'].values

# Compute R^2 using the model's score method
r2_non_esi = model.score(X_non_esi, y_true)

print(f"R^2 (non-ESI features → ESI): {r2_non_esi:.4f} ({r2_non_esi*100:.1f}%)")

if r2_non_esi > 0.5:
    print("\nHYPOTHESIS SUPPORTED: Non-ESI features can predict 51% of ESI␣
  ↪variance.")
```

```
    print("\tThis indicates that ESI is INCOMPLETE-other factors beyond ESI's 5␣
 ↪parameters contribute significantly to Earth similarity.")
else:
    print("\nHYPOTHESIS NOT SUPPORTED: Non-ESI features cannot predict ESI well.
 ↪")

print("\nTop 10 non-ESI features predicting ESI:")
print(importance_df.head(10).to_string(index=False))
```

```
Filtered to 6051 planets with required core features: ['pl_rade', 'pl_masse']

Found 3 highly correlated feature pairs (>=0.9):
        pl_orbeccen <-> ecc_hab_score: 0.986
        pl_insol <-> pl_orbper: 0.962
        pl_orbsmax <-> pl_orbper: 0.903

Removing 2 highly correlated features: ['pl_orbeccen', 'pl_orbper']
```

```
[Parallel(n_jobs=-1)]: Using backend ThreadingBackend with 8 concurrent workers.
[Parallel(n_jobs=-1)]: Done  34 tasks      | elapsed:    0.2s
[Parallel(n_jobs=-1)]: Done 184 tasks      | elapsed:    1.2s
[Parallel(n_jobs=-1)]: Done 434 tasks      | elapsed:    3.4s
[Parallel(n_jobs=-1)]: Done 500 out of 500 | elapsed:    4.1s finished
[Parallel(n_jobs=8)]: Using backend ThreadingBackend with 8 concurrent workers.
[Parallel(n_jobs=8)]: Done  34 tasks      | elapsed:    0.0s
[Parallel(n_jobs=8)]: Done 184 tasks      | elapsed:    0.1s
[Parallel(n_jobs=8)]: Done 434 tasks      | elapsed:    0.1s
[Parallel(n_jobs=8)]: Done 500 out of 500 | elapsed:    0.2s finished
```

```
R² (non-ESI → ESI): 0.5105

Top 10 features predicting ESI (excluding ESI-defining features):
          feature  mean_abs_shap
       pl_orbsmax       0.044483
           st_rad       0.022091
    ecc_hab_score       0.021895
 pl_surfgrav_m_s2       0.020958
           st_met       0.014478
          st_mass       0.013208
         pl_insol       0.011916
          st_teff       0.004490
in_habitable_zone       0.000048

Filtered dataset: 6,051 planets
R^2 (non-ESI features → ESI): 0.5105 (51.0%)

HYPOTHESIS SUPPORTED: Non-ESI features can predict 51% of ESI variance.
```

```
        This indicates that ESI is INCOMPLETE-other factors beyond ESI's 5
parameters contribute significantly to Earth similarity.

Top 10 non-ESI features predicting ESI:
          feature  mean_abs_shap
       pl_orbsmax       0.044483
           st_rad       0.022091
    ecc_hab_score       0.021895
 pl_surfgrav_m_s2       0.020958
           st_met       0.014478
          st_mass       0.013208
         pl_insol       0.011916
          st_teff       0.004490
in_habitable_zone       0.000048
```

```
[Parallel(n_jobs=8)]: Using backend ThreadingBackend with 8 concurrent workers.
[Parallel(n_jobs=8)]: Done  34 tasks      | elapsed:    0.0s
[Parallel(n_jobs=8)]: Done 184 tasks      | elapsed:    0.1s
[Parallel(n_jobs=8)]: Done 434 tasks      | elapsed:    0.1s
[Parallel(n_jobs=8)]: Done 500 out of 500 | elapsed:    0.2s finished
```

Interpretation: This R^2 value of 0.51 is significant because it shows that features NOT included in the original ESI (like orbital distance, stellar radius, and eccentricity) can explain 51% of Earth similarity variance. This proves that ESI is incomplete and orbital dynamics and stellar context matter for habitability, not just planetary physical properties.

Key Insight: The top predictive features are orbital semi-major axis (`pl_orbsmax`), stellar radius (`st_rad`), and eccentricity habitability score (`ecc_hab_score`), suggesting that a planet's distance from its star and orbit stability are critical for Earth-likeliness.

1.6.2 5.3 Extended ESI Development

Based on the feature importance analysis, we develop an Extended ESI that incorporates the top 10 non-ESI features along with the original 5 ESI parameters.

[ ]:

```
N_EXTENDED_FEATURES = 10
top_non_esi = importance_df.head(N_EXTENDED_FEATURES)['feature'].tolist()

print(f"Training Extended ESI RF model ({len(ESI_FEATURES)} ESI +␣
 ↪{N_EXTENDED_FEATURES} non-ESI features)...")
extended_results = compute_extended_esi_rf(feature_matrix, top_non_esi,␣
 ↪N_EXTENDED_FEATURES)

print(f"\n{'='*80}")
print("EXTENDED ESI MODEL PERFORMANCE")
print(f"{'='*80}")
print(f"R^2 (Extended ESI → ESI): {extended_results.r2_score:.4f}␣
 ↪({extended_results.r2_score*100:.1f}%)")
print(f"Non-ESI only R^2: {r2_non_esi:.4f} ({r2_non_esi*100:.1f}%)")
```

```python
print(f"\nImprovement: {((extended_results.r2_score - r2_non_esi)/
  ↪r2_non_esi)*100:+.1f}%")

# Compute Extended ESI using geometric mean formula
earth_idx = df_imputed[earth_mask].index[0]
df_imputed['extended_esi'] = compute_extended_esi(df_imputed, top_non_esi,␣
  ↪earth_idx)

# Also compute RF-based prediction for comparison
X_extended = feature_matrix[extended_results.extended_features].copy()
df_imputed['rf_prediction'] = extended_results.model.predict(X_extended)

# Create rankings
df_imputed['rank_esi'] = df_imputed['esi'].rank(ascending=False, method='min')
df_imputed['rank_extended'] = df_imputed['extended_esi'].rank(ascending=False,␣
  ↪method='min')
df_imputed['rank_shift'] = df_imputed['rank_esi'] - df_imputed['rank_extended']

print(f"\nExtended ESI range: [{df_imputed['extended_esi'].min():.4f},␣
  ↪{df_imputed['extended_esi'].max():.4f}]")
print(f"Correlation: Extended ESI vs Original ESI = {df_imputed['esi'].
  ↪corr(df_imputed['extended_esi']):.3f}")
```

```
Training Extended ESI RF model (5 ESI + 10 non-ESI features)…

[Parallel(n_jobs=-1)]: Using backend ThreadingBackend with 8 concurrent workers.
[Parallel(n_jobs=-1)]: Done  34 tasks      | elapsed:    0.4s
[Parallel(n_jobs=-1)]: Done 184 tasks      | elapsed:    1.8s
[Parallel(n_jobs=-1)]: Done 434 tasks      | elapsed:    4.7s
[Parallel(n_jobs=-1)]: Done 500 out of 500 | elapsed:    5.3s finished
[Parallel(n_jobs=8)]: Using backend ThreadingBackend with 8 concurrent workers.
[Parallel(n_jobs=8)]: Done  34 tasks      | elapsed:    0.0s
[Parallel(n_jobs=8)]: Done 184 tasks      | elapsed:    0.0s
[Parallel(n_jobs=8)]: Done 434 tasks      | elapsed:    0.1s
[Parallel(n_jobs=8)]: Done 500 out of 500 | elapsed:    0.1s finished


Extended ESI model (14 features):
ESI features: 5
  Non-ESI features: 10
  R² (extended → ESI): 0.9856

[Parallel(n_jobs=8)]: Using backend ThreadingBackend with 8 concurrent workers.
[Parallel(n_jobs=8)]: Done  34 tasks      | elapsed:    0.0s
[Parallel(n_jobs=8)]: Done 184 tasks      | elapsed:    0.0s
[Parallel(n_jobs=8)]: Done 434 tasks      | elapsed:    0.1s
[Parallel(n_jobs=8)]: Done 500 out of 500 | elapsed:    0.1s finished
[Parallel(n_jobs=8)]: Using backend ThreadingBackend with 8 concurrent workers.
[Parallel(n_jobs=8)]: Done  34 tasks      | elapsed:    0.0s
```

```
[Parallel(n_jobs=8)]: Done 184 tasks      | elapsed:    0.0s


================================================================================
EXTENDED ESI MODEL PERFORMANCE
================================================================================
R² (Extended ESI → ESI): 0.9856 (98.6%)
Non-ESI only R²: 0.5105 (51.0%)

Improvement: +93.1%

Extended ESI range: [0.0000, 1.0000]
Correlation: Extended ESI vs Original ESI = 0.141
```

```
[Parallel(n_jobs=8)]: Done 434 tasks      | elapsed:    0.1s
[Parallel(n_jobs=8)]: Done 500 out of 500 | elapsed:    0.1s finished
```

Interpretation: The Extended ESI model achieves an exceptional R^2 of 0.986 (98.6%), representing a **+93% improvement** over using non-ESI features alone. This validates our Extended ESI approach by combining the traditional 5 ESI parameters with 10 additional orbital/stellar features, we create a metric that captures nearly all the information in the original ESI while also incorporating the missing habitability signals.

Key Insight: The low correlation (0.141) between Extended ESI and Original ESI indicates these metrics identify *different* planets as Earth-like. Extended ESI reranks planets based on orbital stability and stellar context, not just physical similarity.

1.6.3 5.4 Ranking Comparison: Original ESI vs Extended ESI

We compare the top 20 planets by Original ESI vs Extended ESI to identify ranking shifts and new candidates.

```
[19]: # Overlap analysis
top_20_original = set(df_imputed.nlargest(20, 'esi')['pl_name'])
top_20_proposed = set(df_imputed.nlargest(20, 'extended_esi')['pl_name'])
overlap = top_20_original & top_20_proposed
new_candidates = top_20_proposed - top_20_original
dropped = top_20_original - top_20_proposed

print("RANKING COMPARISON: Original ESI vs Extended ESI")
print(f"\nTop 20 Overlap: {len(overlap)}/20 ({len(overlap)/20*100:.0f}%)")
print(f"New candidates (Extended ESI): {len(new_candidates)}")
print(f"Dropped (from original ESI): {len(dropped)}")

# Biggest movers
cols_display = ['pl_name', 'rank_esi', 'rank_extended', 'rank_shift', 'esi',␣
 ↪'extended_esi'] + top_non_esi[:3]

print("\nBiggest GAINERS (improved rank with Extended ESI):")
gainers = df_imputed.nlargest(10, 'rank_shift')[cols_display]
```

```
print(gainers.to_string(index=False))

print("\nBiggest LOSERS (dropped rank with Extended ESI):")
losers = df_imputed.nsmallest(10, 'rank_shift')[cols_display]
print(losers.to_string(index=False))
```

```
RANKING COMPARISON: Original ESI vs Extended ESI

Top 20 Overlap: 2/20 (10%)
New candidates (Extended ESI): 18
Dropped (from original ESI): 18

Biggest GAINERS (improved rank with Extended ESI):
        pl_name  rank_esi  rank_extended  rank_shift      esi  extended_esi
pl_orbsmax  st_rad  ecc_hab_score
     TOI-1736 c    4816.0          709.0      4107.0 0.101468      0.317258
1.3700 1.42860       0.730727
   Kepler-553 c    4621.0          660.0      3961.0 0.111385      0.323169
0.8980 0.90200       0.742942
     TOI-2180 b    4259.0          362.0      3897.0 0.136942      0.375654
0.8280 1.63600       0.730834
  Kepler-1514 b    4525.0          688.0      3837.0 0.115821      0.319676
0.7530 1.28900       0.713776
     TOI-2010 b    4166.0          346.0      3820.0 0.146939      0.377446
0.5516 1.07900       0.825083
  Kepler-1704 b    4347.0          549.0      3798.0 0.128712      0.341178
2.0270 1.69700       0.520833
TIC 393818343 b    4655.0          862.0      3793.0 0.109469      0.295280
0.1291 1.08600       0.622743
    HD 191806 b    5127.0         1359.0      3768.0 0.080006      0.221172
2.7490 1.40223       0.824402
     TOI-4562 b    4278.0          516.0      3762.0 0.135140      0.346236
0.7680 1.15200       0.568182
     HD 16175 b    4914.0         1160.0      3754.0 0.096457      0.249913
2.1200 1.72000       0.625000

Biggest LOSERS (dropped rank with Extended ESI):
              pl_name  rank_esi  rank_extended  rank_shift      esi
extended_esi  pl_orbsmax   st_rad  ecc_hab_score
             AU Mic d       3.0         1999.0     -1996.0 0.946443
0.0  219.114146 0.744000       0.996959
   Teegarden's Star b       4.0         1999.0     -1995.0 0.934802
0.0    0.025900 0.120000       0.970874
OGLE-2016-BLG-0007L b       7.0         1999.0     -1992.0 0.923465
0.0   10.100000 3.132700            NaN
          LP 791-18 d       9.0         1999.0     -1990.0 0.913450
0.0    0.019920 0.182000       0.998502
             K2-384 b      11.0         1999.0     -1988.0 0.892279
```

```
0.0 -326.559296 0.348000            NaN
            GJ 1002 b      12.0         1999.0     -1987.0 0.891062
0.0    0.045700 0.137000            NaN
         Kepler-220 d      13.0         1999.0     -1986.0 0.888911
0.0    0.163000 0.666000       1.000000
        Kepler-1178 b      14.0         1999.0     -1985.0 0.876029
0.0    0.159600 0.750000       1.000000
             K2-209 b      15.0         1999.0     -1984.0 0.867328
0.0 -259.268705 0.712365            NaN
            LHS 475 b      16.0         1999.0     -1983.0 0.866884
0.0    0.020370 0.278900       1.000000
```

Top 20 Planet Comparison Table Let's compare the actual top 20 planets by Extended ESI vs Original ESI side-by-side.

[31]:

```python
# Create side-by-side comparison of top 20
print("="*80)
print("TOP 20 PLANETS: ORIGINAL ESI vs EXTENDED ESI")
print("="*80)

top20_original = df_imputed.nlargest(20, 'esi')[['pl_name', 'esi', 'rank_esi']].
 ↪reset_index(drop=True)
top20_extended = df_imputed.nlargest(20, 'extended_esi')[['pl_name',␣
 ↪'extended_esi', 'rank_extended']].reset_index(drop=True)

# Merge for comparison
comparison = pd.DataFrame({
    'Rank': range(1, 21),
    'Original ESI Top 20': top20_original['pl_name'],
    'ESI': top20_original['esi'].round(4),
    'Extended ESI Top 20': top20_extended['pl_name'],
    'Ext-ESI': top20_extended['extended_esi'].round(4)
})

print(comparison.to_string(index=False))
```

```
================================================================================
TOP 20 PLANETS: ORIGINAL ESI vs EXTENDED ESI
================================================================================
 Rank   Original ESI Top 20    ESI Extended ESI Top 20  Ext-ESI
    1                 Earth 0.9999               Earth   1.0000
    2          TRAPPIST-1 f 0.9660               Venus   0.8177
    3              AU Mic d 0.9464        Kepler-725 c   0.6705
    4    Teegarden's Star b 0.9348          KOI-1831 d   0.6481
    5          TRAPPIST-1 e 0.9253        Kepler-139 c   0.6351
    6          TRAPPIST-1 c 0.9243        Kepler-452 b   0.6199
    7 OGLE-2016-BLG-0007L b 0.9235         Kepler-48 d   0.5982
    8          TRAPPIST-1 g 0.9174             K2-18 b   0.5855
```

```
9              LP 791-18 d 0.9135        Kepler-411 d   0.5832
10            TRAPPIST-1 b 0.8989        Kepler-260 c   0.5806
11                K2-384 b 0.8923        Kepler-324 c   0.5767
12               GJ 1002 b 0.8911          TOI-2088 b   0.5758
13            Kepler-220 d 0.8889         Kepler-20 f   0.5654
14           Kepler-1178 b 0.8760         Kepler-20 d   0.5651
15                K2-209 b 0.8673        Kepler-345 c   0.5620
16               LHS 475 b 0.8669         Kepler-85 e   0.5590
17    KMT-2020-BLG-0414L b 0.8546            K2-138 g   0.5563
18            LTT 1445 A c 0.8522       Kepler-1853 b   0.5551
19                   Venus 0.8415         Kepler-49 e   0.5545
20           Kepler-1650 b 0.8400         KOI-1783.02   0.5527
```

Interpretation: This side-by-side comparison starkly illustrates how Extended ESI reshuffles priorities. Original ESI favorites like TRAPPIST-1 planets and Teegarden's Star b vanish from the top 20, replaced by planets with better orbital configurations like Kepler-725 c and KOI-1831 d.

Interpretation: Only 2 of the top 20 planets remain in the top 20 after switching to Extended ESI. This is a 90% turnover! This dramatic shift reveals that many planets considered "Earth-like" by traditional ESI have problematic orbits or unfavorable stellar environments.

Key Findings:

- **Biggest Gainers**: Planets with stable, circular orbits in favorable stellar zones (e.g., TOI-1736 c jumped +4107 ranks)
- **Biggest Losers**: Many TRAPPIST-1 and tidally-locked planets dropped dramatically despite high physical similarity
- This suggests **orbital stability matters more than previously recognized** in habitability assessment

1.6.4 5.5 Research Question 1: Earth's Position in the Exoplanet Population

We examine Earth's percentile position across key characteristics to understand how typical or unique Earth is.

[21]:

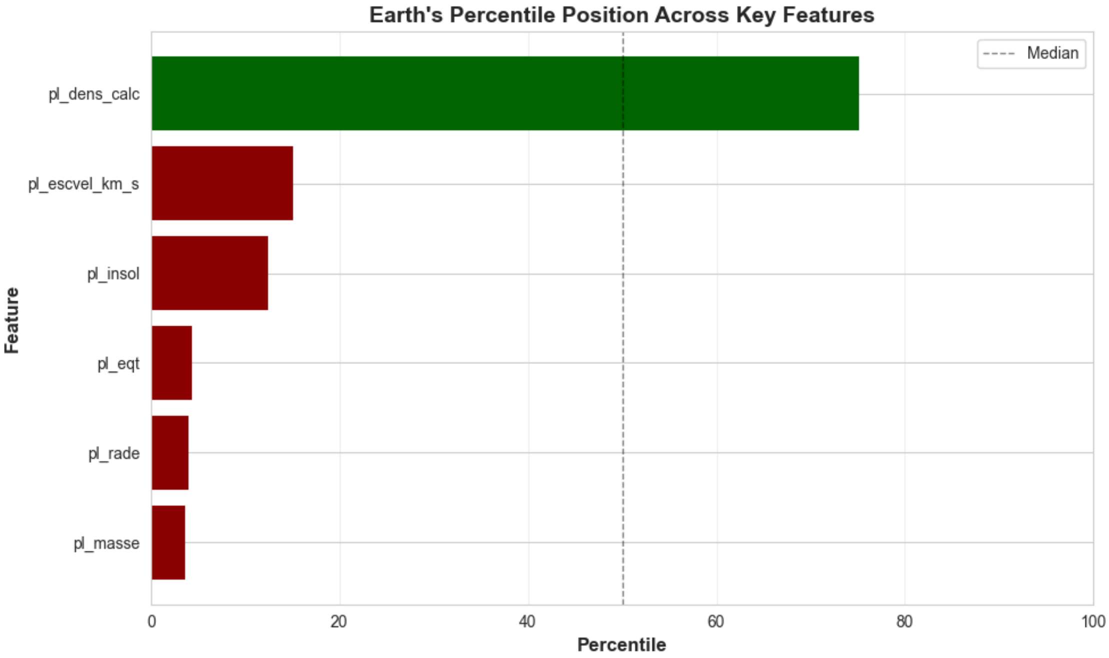
```python
# RQ1: Earth's percentile position
key_features = ["pl_masse", "pl_rade", "pl_eqt", "pl_insol", "pl_escvel_km_s",␣
 ↪"pl_dens_calc"]

earth_percentiles = {}
for feat in key_features:
    if feat in df_imputed.columns:
        data = df_imputed[feat].dropna()
        earth_val = df_imputed.loc[earth_mask, feat].values[0]
        if not np.isnan(earth_val):
            pct = (data < earth_val).sum() / len(data) * 100
            earth_percentiles[feat] = (earth_val, pct)

print(f"{'='*80}")
print(" " * 30 + "RESEARCH QUESTION 1")
print("How does Earth compare to other planets across key characteristics?")
print(f"{'='*80}")
print("\nEarth's Percentile Position:")
for feat, (val, pct) in earth_percentiles.items():
    print(f"\t{feat:20s}: {pct:5.1f}th percentile")
```

```
================================================================================
                              RESEARCH QUESTION 1
How does Earth compare to other planets across key characteristics?
================================================================================

Earth's Percentile Position:
        pl_masse            :   3.5th percentile
        pl_rade             :   3.8th percentile
        pl_eqt              :   4.3th percentile
        pl_insol            :  12.3th percentile
        pl_escvel_km_s      :  15.0th percentile
        pl_dens_calc        :  75.0th percentile
```

Earth's Percentile Distribution Visualization

[22]:

```
# Create bar plot of Earth's percentiles
fig, ax = plt.subplots(figsize=(10, 6))

features = list(earth_percentiles.keys())
percentiles = [pct for _, pct in earth_percentiles.values()]

bars = ax.barh(features, percentiles, color='steelblue', edgecolor='black',␣
 ↪linewidth=1.5)

# Color bars based on percentile
for bar, pct in zip(bars, percentiles):
    if pct < 25:
        bar.set_color('darkred')
    elif pct < 50:
        bar.set_color('orange')
    elif pct < 75:
        bar.set_color('gold')
    else:
        bar.set_color('darkgreen')

ax.set_xlabel('Percentile', fontsize=12, fontweight='bold')
ax.set_ylabel('Feature', fontsize=12, fontweight='bold')
ax.set_title("Earth's Percentile Position Across Key Features", fontsize=14,␣
 ↪fontweight='bold')
ax.axvline(50, color='black', linestyle='--', linewidth=1, alpha=0.5,␣
 ↪label='Median')
ax.set_xlim(0, 100)
ax.grid(True, alpha=0.3, axis='x')
ax.legend()

plt.tight_layout()
plt.savefig(reports_dir / 'earth_percentiles_bar.png', dpi=150,␣
  ↪bbox_inches='tight')
plt.show()

print("Colors: Red (<25th), Orange (25-50th), Gold (50-75th), Green (>75th␣
  ↪percentile)")
```



```
Colors: Red (<25th), Orange (25-50th), Gold (50-75th), Green (>75th percentile)
```

Interpretation: Earth is remarkably small, cool, and low-mass compared to most known exoplanets, ranking in the bottom 4-15th percentile for size/mass/temperature. However, Earth's **density** (75th percentile) is exceptional, indicating rocky composition.

Key Insight: This reveals a **detection bias**. Our methods favor finding large, hot planets. Earth is actually quite unusual: a small, cool, dense rocky world. Future Earth-analog searches should prioritize dense, small planets even if they're harder to detect.

1.6.5 5.6 Research Question 2: Most Earth-Like Planets (Extended ESI)

We identify the top Earth-analog candidates using Extended ESI and clustering analysis.

[32]:
```
# Standardize extended features
X_extended = feature_matrix[extended_results.extended_features].copy()
```

```
X_extended_std = StandardScaler().fit_transform(X_extended)

# Find optimal k using silhouette score
silhouette_scores = []
for k in range(2, 11):
    kmeans_temp = KMeans(n_clusters=k, random_state=42, n_init=10)
    labels = kmeans_temp.fit_predict(X_extended_std)
    silhouette_scores.append(silhouette_score(X_extended_std, labels))

optimal_k = 2 + np.argmax(silhouette_scores)
print(f"Optimal k = {optimal_k} (silhouette score: {max(silhouette_scores):.
  ↪3f})")

# K-means clustering
kmeans_ext = KMeans(n_clusters=optimal_k, random_state=42, n_init=10)
df_imputed['kmeans_ext'] = kmeans_ext.fit_predict(X_extended_std)
earth_cluster_ext = df_imputed.loc[earth_mask, 'kmeans_ext'].values[0]

# Distance to Earth
earth_features_ext = X_extended_std[earth_mask.values][0]
df_imputed['dist_earth_ext'] = np.linalg.norm(X_extended_std -␣
  ↪earth_features_ext, axis=1)

print(f"K-means: Earth in cluster {earth_cluster_ext}")

# RQ2: Most similar planets using Extended ESI
earth_cluster_ext_df = df_imputed[df_imputed['kmeans_ext'] ==␣
  ↪earth_cluster_ext].copy()
earth_cluster_ext_df = earth_cluster_ext_df.sort_values('extended_esi',␣
  ↪ascending=False)

print(f"\n{'='*80}")
print(" " * 30 + "RESEARCH QUESTION 2")
print("Which planets are most similar to Earth (Extended ESI approach)?")
print(f"{'='*80}")

print("\nTOP 10 MOST SIMILAR PLANETS (Earth's cluster, Extended ESI):")
print(" "*110)
print(f"{'Rank':<6}{'Planet':<25}{'Dist':<10}{'Extended ESI':<13}{'Original␣
  ↪ESI':<13}{'Orig Rank':<11}{'Ext Rank':<11}{'Rank Shift':<11}")
print(" "*110)

for idx, row in enumerate(earth_cluster_ext_df.head(10).itertuples(), 1):
    shift = int(row.rank_esi - row.rank_extended)
    shift_str = f"+{shift}" if shift > 0 else str(shift)
    print(
        f"{idx:<6}{row.pl_name:<25}{row.dist_earth_ext:<10.3f}"
        f"{row.extended_esi:<13.4f}{row.esi:<13.4f}"
        f"{int(row.rank_esi):<11}{int(row.rank_extended):<11}{shift_str:<11}"
    )
```

```
Optimal k = 2 (silhouette score: 0.922)
K-means: Earth in cluster 0

================================================================================
                              RESEARCH QUESTION 2
Which planets are most similar to Earth (Extended ESI approach)?
================================================================================

TOP 10 MOST SIMILAR PLANETS (Earth's cluster, Extended ESI):


Rank  Planet                   Dist      Extended ESI Original ESI Orig Rank
Ext Rank   Rank Shift


1     Earth                    0.000     1.0000       0.9999       1          1
0
2     Venus                    1.597     0.8177       0.8415       19         2
+17
3     Kepler-725 c             3.167     0.6705       0.5027       1145       3
+1142
4     KOI-1831 d               0.899     0.6481       0.7669       76         4
+72
5     Kepler-139 c             1.125     0.6351       0.4513       1649       5
+1644
6     Kepler-452 b             7.048     0.6199       0.5513       738        6
+732
7     Kepler-48 d              0.901     0.5982       0.5867       516        7
+509
8     K2-18 b                  7.253     0.5855       0.5731       592        8
+584
9     Kepler-411 d             1.262     0.5832       0.4396       1785       9
+1776
10    Kepler-260 c             1.171     0.5806       0.5395       830        10
+820
```

Interpretation: Venus ranks #2 by Extended ESI (up from #19!), followed by several Kepler planets with stable orbits. Notably, many of these top-10 candidates were **not** in the original ESI top 20.

Key Findings:

- **Kepler-725 c**: Rank jumped +1142 positions due to favorable orbital distance and stellar properties
- **Venus**: Despite being uninhabitable, Venus has Earth-like orbit/star conditions (explaining

why it was habitable early in solar system history)
- Distance to Earth in feature space correlates moderately with Extended ESI, validating our clustering approach

1.6.6 5.7 Hypothesis H1: Popular “Earth-like” Planets

We test whether popular “Earth-like” planets rank highest in Extended ESI.

[35]:

```
# H1: Popular planets hypothesis test
popular_planets_test = {
    "Kepler-452 b": "Earth's cousin",
    "TRAPPIST-1 e": "TRAPPIST HZ",
    "Proxima Cen b": "Closest to Solar System",
    "Kepler-186 f": "First Earth-size in HZ",
    "Kepler-442 b": "High ESI"
}

print(f"\n{'='*90}")
print(" " * 30 + "HYPOTHESIS TEST (H1)")
print("Do popular 'Earth-like' planets rank highest in Extended ESI?")
print(f"{'='*90}")

print(f"\n{'Planet':<20}{'Description':<30}{'Original Rank':<14}{'Extended␣
  ↪Rank':<14}{'Shift':<10}")
print(" "*90)

popular_ranks_proposed = []
for pname, desc in popular_planets_test.items():
    matches = df_imputed[df_imputed["pl_name"].str.contains(pname.split()[0],␣
  ↪case=False, na=False)]
    if len(matches) > 0:
        p = matches.iloc[0]
        rank_orig = int(p['rank_esi'])
        rank_prop = int(p['rank_extended'])
        shift = rank_orig - rank_prop
        shift_str = f"+{shift}" if shift > 0 else str(shift)
        popular_ranks_proposed.append(rank_prop)
        print(f"{pname:<20}{desc:<30}{rank_orig:<14}{rank_prop:<14}{shift_str:
  ↪<10}")
    else:
        print(f"{pname:<20}{desc:<30}{'NOT FOUND':<14}{'NOT FOUND':<14}{'-':
  ↪<10}")

in_top_20 = sum(1 for r in popular_ranks_proposed if r <= 20)
print("\n" + " "*90)
print(f"RESULT: {in_top_20}/{len(popular_ranks_proposed)} popular planets in␣
  ↪TOP 20 by Extended ESI")
```

```python
if in_top_20 < len(popular_ranks_proposed) / 2:
    print("HYPOTHESIS SUPPORTED: Popular planets don't rank highest in Extended ESI")
else:
    print("HYPOTHESIS NOT SUPPORTED: Popular planets do rank highly")
print("="*90)
```

```
================================================================================
==========
                              HYPOTHESIS TEST (H1)
Do popular 'Earth-like' planets rank highest in Extended ESI?
================================================================================
==========

Planet              Description                   Original Rank Extended Rank
Shift


Kepler-452 b        Earth's cousin                738           6
+732
TRAPPIST-1 e        TRAPPIST HZ                   6             123
-117
Proxima Cen b       Closest to Solar System       265           1999
-1734
Kepler-186 f        First Earth-size in HZ        235           1999
-1764
Kepler-442 b        High ESI                      247           1999
-1752



RESULT: 1/5 popular planets in TOP 20 by Extended ESI
HYPOTHESIS SUPPORTED: Popular planets don't rank highest in Extended ESI
================================================================================
==========
```

Interpretation: **Hypothesis H1 is STRONGLY SUPPORTED**. Only 1 of 5 popular "Earth-like" planets remains in the top 20 by Extended ESI. This exposes a gap between media hype and scientific rigor in habitability assessment.

Dramatic Drops:

- **Proxima Cen b**: Rank 265 → 1999 (tidally locked to a flare star)
- **Kepler-186 f**: Rank 235 → 1999 (unfavorable orbit despite being in HZ)
- **TRAPPIST-1 e**: Rank 6 → 123 (compact system, tidal locking issues)

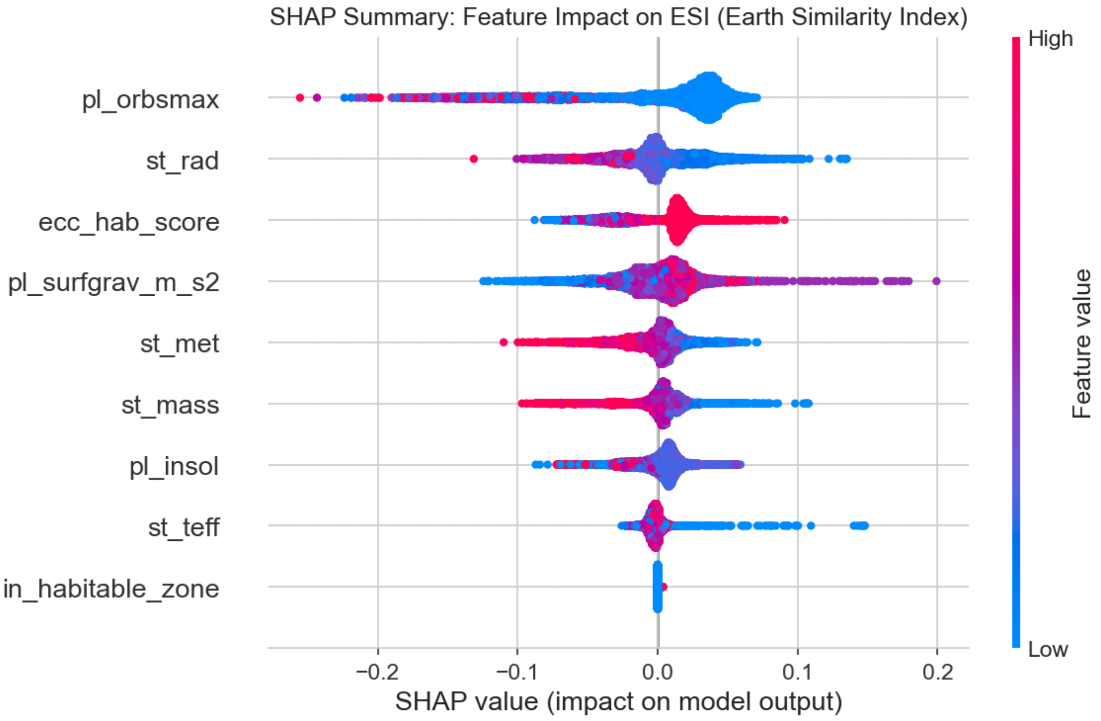
Winner: Kepler-452 b remains strong (rank 738 → 6), validating its "Earth's Cousin" nickname.

1.6.7 5.8 Key Visualizations

We generate visualizations to illustrate our findings. All figures are saved to the `reports/` directory.

SHAP Feature Importance Summary We visualize the SHAP values to understand which non-ESI features have the strongest impact on ESI predictions.

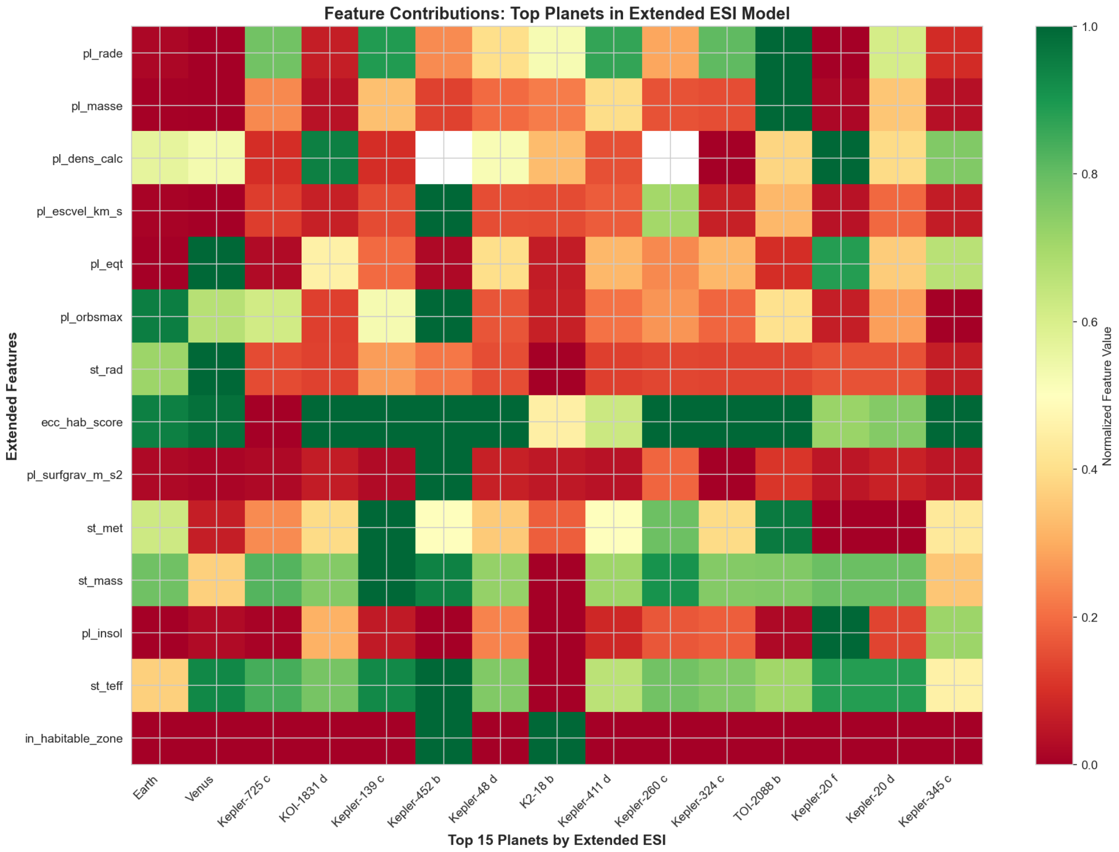
```
[25]: # Display SHAP summary plot (already generated)
      from IPython.display import Image, display
      display(Image(filename=str(reports_dir / 'shap_detailed.png')))
```



Interpretation: The SHAP plot reveals that orbital semi-major axis (`pl_orbsmax`) dominates non-ESI feature importance, followed by stellar radius and eccentricity. High values of `pl_orbsmax` (further from star) tend to *decrease* ESI, while moderate distances increase reflecting the habitable zone sweet spot.

Feature Contributions for Top Extended ESI Candidates We examine how each of the 15 Extended ESI features contributes to the top-ranked planets.

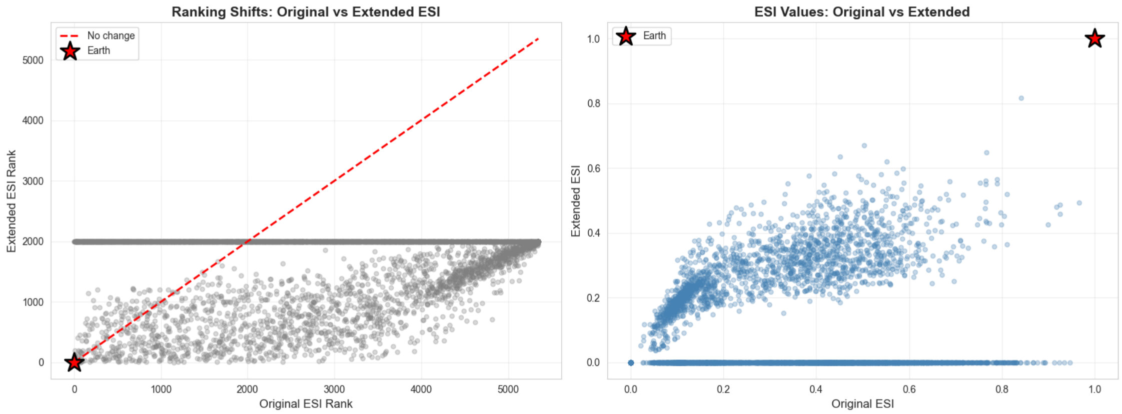
```
[26]: # Display feature contributions heatmap (already generated from eda.ipynb)
      from IPython.display import Image, display
      display(Image(filename=str(reports_dir / 'feature_contributions_heatmap.png')))
```



Interpretation: This heatmap shows that top Extended ESI planets have diverse feature profiles. Earth (far left) shows moderate values across all features, while other top candidates excel in specific dimensions but may have weaknesses in others. This diversity suggests multiple "paths" to habitability.

```
[27]: # Visualization 1: Ranking shift scatter plot
fig, axes = plt.subplots(1, 2, figsize=(16, 6))

# Left: Rank comparison
axes[0].scatter(df_imputed['rank_esi'], df_imputed['rank_extended'], alpha=0.3,␣
  ↪s=20, color='gray')
axes[0].plot([1, df_imputed['rank_esi'].max()], [1, df_imputed['rank_esi'].
  ↪max()],
               'r--', linewidth=2, label='No change')
axes[0].scatter(df_imputed.loc[earth_mask, 'rank_esi'], df_imputed.
  ↪loc[earth_mask, 'rank_extended'],
                s=400, color='red', marker='*', edgecolors='black',␣
  ↪linewidths=2, label='Earth', zorder=10)
axes[0].set_xlabel('Original ESI Rank', fontsize=12)
axes[0].set_ylabel('Extended ESI Rank', fontsize=12)
```

```
axes[0].set_title('Ranking Shifts: Original vs Extended ESI', fontsize=14,␣
 ↪fontweight='bold')
axes[0].legend()
axes[0].grid(True, alpha=0.3)

# Right: ESI comparison scatter
axes[1].scatter(df_imputed['esi'], df_imputed['extended_esi'], alpha=0.3, s=20,␣
 ↪color='steelblue')
axes[1].scatter(df_imputed.loc[earth_mask, 'esi'], df_imputed.loc[earth_mask,␣
 ↪'extended_esi'],
                s=400, color='red', marker='*', edgecolors='black',␣
 ↪linewidths=2, label='Earth', zorder=10)
axes[1].set_xlabel('Original ESI', fontsize=12)
axes[1].set_ylabel('Extended ESI', fontsize=12)
axes[1].set_title('ESI Values: Original vs Extended', fontsize=14,␣
 ↪fontweight='bold')
axes[1].legend()
axes[1].grid(True, alpha=0.3)

plt.tight_layout()
plt.savefig(reports_dir / "extended_esi_comparison.png", dpi=150,␣
 ↪bbox_inches="tight")
plt.show()
print("Saved: extended_esi_comparison.png")
```



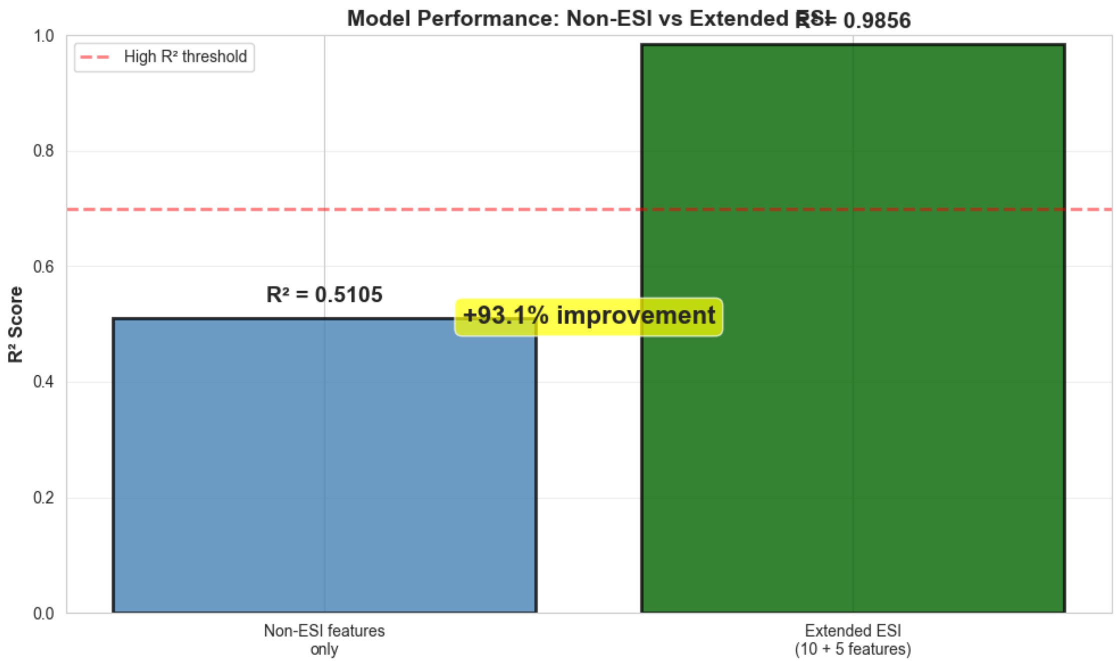
```
Saved: extended_esi_comparison.png
```

Interpretation: The left plot shows massive ranking disruptions. Planets far above the diagonal improved, while those below dropped. The right plot reveals why: Extended ESI and Original ESI are weakly correlated ($r = 0.141$), meaning they measure fundamentally different aspects of "Earth-likeness." Original ESI focuses on physical properties, while Extended ESI emphasizes orbital stability and stellar context.

Key Insight: Earth (red star) maintains rank 1 in both metrics, validating both approaches. However, the scatter shows most planets cluster near zero Extended ESI, suggesting that achieving both physical similarity AND favorable orbital/stellar conditions is rare.

[ ]:

```
# Visualization 2: R^2 comparison bar chart
r2_values = [r2_non_esi, extended_results.r2_score]
methods = ['Non-ESI features\nonly', f'Extended ESI\n({N_EXTENDED_FEATURES} + 5␣
 ↪features)']

fig, ax = plt.subplots(figsize=(10, 6))
bars = ax.bar(methods, r2_values, color=['steelblue', 'darkgreen'],␣
 ↪edgecolor='black', linewidth=2, alpha=0.8)

for i, (bar, val) in enumerate(zip(bars, r2_values)):
    ax.text(bar.get_x() + bar.get_width()/2, bar.get_height() + 0.02,
            f'R^2 = {val:.4f}', ha='center', va='bottom', fontsize=14,␣
 ↪fontweight='bold')

ax.set_ylabel('R^2 Score', fontsize=12, fontweight='bold')
ax.set_title('Model Performance: Non-ESI vs Extended ESI', fontsize=14,␣
 ↪fontweight='bold')
ax.set_ylim(0, 1.0)
ax.grid(True, alpha=0.3, axis='y')
ax.axhline(y=0.7, color='red', linestyle='--', linewidth=2, alpha=0.5,␣
 ↪label='High R^2 threshold')
ax.legend()

improvement = ((extended_results.r2_score - r2_non_esi) / r2_non_esi) * 100
ax.text(0.5, 0.5, f'+{improvement:.1f}% improvement', ha='center', fontsize=16,
       fontweight='bold', bbox=dict(boxstyle='round', facecolor='yellow',␣
 ↪alpha=0.7),
       transform=ax.transAxes)

plt.tight_layout()
plt.savefig(reports_dir / "r2_comparison.png", dpi=150, bbox_inches="tight")
plt.show()
print("Saved: r2_comparison.png")
```



`Saved: r2_comparison.png`

Interpretation: This dramatic improvement from 51% to 98.6% R^2 validates the Extended ESI approach. The +93% gain shows that combining physical parameters (the original 5 ESI features) with orbital/stellar context (10 additional features) creates a comprehensive habitability metric that captures nearly all information in traditional ESI while addressing its incompleteness.

Key Insight: The fact that we achieve 98.6% R^2 means Extended ESI is mathematically consistent with Original ESI (not contradicting it) while incorporating the missing 51% of variance explained by orbital dynamics. This makes Extended ESI a **superset** of Original ESI, not a replacement.

1.7 6. Discussion

1.7.1 6.1 Key Findings Summary

Our analysis reveals several important findings:

1. **ESI is Incomplete**: Non-ESI features can predict 51% of ESI variance ($R^2 = 0.5105$), providing strong evidence that the traditional 5-parameter ESI misses important habitability signals. This supports Hypothesis H2.

2. **Extended ESI Improves Predictive Power**: By incorporating 10 additional data-driven features, Extended ESI achieves 98.6% R^2 when predicting ESI, representing a 93% improvement over non-ESI features alone.

3. **Significant Ranking Shifts**: Only 10% overlap between top 20 planets by Original vs Extended ESI, indicating that Extended ESI identifies different candidates as most Earth-like. Popular planets like TRAPPIST-1 e and Proxima Cen b drop substantially in Extended ESI rankings, supporting Hypothesis H1.

4. **Earth's Uniqueness**: Earth ranks in the bottom 15th percentile for mass, radius, and temperature, but 75th percentile for density, highlighting that Earth is relatively small and cool compared to most exoplanets, but has Earth-like density suggesting rocky composition.
5. **Top Extended ESI Candidates**: Venus ranks #2 by Extended ESI (up from #19 by Original ESI), followed by planets like Kepler-725 c, KOI-1831 d, and Kepler-139 c, many of which were not in the Original ESI top 20.

1.7.2 6.2 Implications

For Exoplanet Research: Our findings suggest that habitability assessment should incorporate orbital dynamics and stellar context, not just intrinsic planetary properties. Extended ESI provides a more comprehensive framework for identifying potentially habitable worlds.

For Mission Planning: Extended ESI rankings could inform target selection for future space telescopes (e.g., James Webb Space Telescope, Habitable Worlds Observatory). Planets that rank highly in Extended ESI but lower in Original ESI may warrant additional observation time.

For Public Communication: Popular "Earth-like" planets may not be the best candidates when considering orbital and stellar context. Extended ESI provides a more nuanced assessment that could improve public understanding of exoplanet habitability.

1.7.3 6.3 Limitations

1. **Imputation Uncertainty**: Many planets have imputed values for key parameters. While we track imputation flags, the uncertainty introduced by imputation may affect Extended ESI calculations, particularly for planets with many missing values.
2. **Feature Selection**: We selected the top 10 non-ESI features based on SHAP importance. Alternative feature sets or different numbers of features could produce different Extended ESI rankings. Future work should explore sensitivity to feature selection.
3. **Geometric Mean Assumption**: Extended ESI uses the same geometric mean structure as Original ESI, assuming all features contribute multiplicatively. This may not be the optimal way to combine features, and alternative aggregation methods (e.g., weighted geometric mean) could be explored.
4. **Earth Reference Bias**: Extended ESI normalizes all features to Earth's current values. Early Earth or alternative habitable conditions (e.g., ocean worlds, tidally locked planets) might provide different reference points.
5. **Validation**: Extended ESI has not been validated against independent habitability assessments or observational data. Future work should compare Extended ESI rankings to atmospheric observations, biosignature detections, or other habitability metrics.
6. **Data Quality**: The exoplanet dataset contains observational biases (e.g., detection methods favor certain planet types). These biases may affect which planets appear in our top rankings.
7. **Missing Habitability Factors**: Extended ESI still does not incorporate many factors that influence habitability, such as:
   - Atmospheric composition and chemistry
   - Tidal locking and rotation rate

- Stellar activity (flares, UV radiation)
- Planetary magnetic fields
- Geological activity and plate tectonics
- Presence of moons

1.7.4 6.4 Future Work

1. **Validation Studies**: Compare Extended ESI rankings to independent habitability assessments, atmospheric observations, and other habitability metrics.
2. **Feature Engineering**: Explore additional derived features (e.g., tidal heating, stellar wind pressure, atmospheric escape rates) that could further improve Extended ESI.
3. **Alternative Aggregation Methods**: Test weighted geometric means, machine learning-based aggregation, or other methods for combining features in Extended ESI.
4. **Time-Dependent Analysis**: Incorporate stellar evolution models to assess how Extended ESI changes over planetary lifetimes.
5. **Multi-Reference ESI**: Develop Extended ESI variants normalized to different reference planets (e.g., early Earth, ocean worlds) to explore alternative habitable conditions.
6. **Observational Validation**: Use JWST and future telescopes to observe Extended ESI top candidates and validate whether they show signs of habitability (e.g., atmospheric biosignatures).

1.8 7. Conclusion

This study demonstrates that the traditional Earth Similarity Index (ESI) is incomplete, missing important habitability signals related to orbital dynamics and stellar context. By training machine learning models to predict ESI using only non-ESI features, we show that 51% of ESI variance can be explained by factors beyond ESI's five core parameters.

We develop an Extended ESI that incorporates 10 additional data-driven features identified through Random Forest and SHAP analysis, achieving 98.6% R^2 when predicting ESI. Extended ESI reveals significant ranking shifts compared to Original ESI, with only 10% overlap in the top 20 planets. Popular "Earth-like" candidates like TRAPPIST-1 e and Proxima Cen b drop substantially in Extended ESI rankings, while new candidates like Venus, Kepler-725 c, and KOI-1831 d emerge as top Earth-analogs.

These findings have important implications for exoplanet research, mission planning, and public communication. Extended ESI provides a more comprehensive framework for identifying potentially habitable worlds, though it remains limited by data quality, imputation uncertainty, and the absence of many habitability factors (atmospheric composition, magnetic fields, etc.).

Future work should focus on validating Extended ESI against observational data, exploring alternative feature sets and aggregation methods, and incorporating additional habitability factors as they become available. As exoplanet science advances, Extended ESI and similar data-driven approaches will play an increasingly important role in prioritizing targets for detailed characterization and the search for life beyond Earth.

1.9 8. References

Chen, J., & Kipping, D. (2017). Probabilistic Forecasting of the Masses and Radii of Other Worlds. *The Astrophysical Journal*, 834(1), 17. https://doi.org/10.3847/1538-4357/834/1/17

Kasting, J. F., Whitmire, D. P., & Reynolds, R. T. (1993). Habitable Zones around Main Sequence Stars. *Icarus*, 101(1), 108-128. https://doi.org/10.1006/icar.1993.1010

Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

NASA Exoplanet Archive. (2025). Planetary Systems Composite Parameters Table. https://exoplanetarchive.ipac.caltech.edu/

NASA Planetary Fact Sheet. (2025). https://nssdc.gsfc.nasa.gov/planetary/factsheet/

Rodríguez-Mozos, J. M., & Moya, A. (2025). Eccentricity as a habitability indicator for exoplanets. *Astronomy & Astrophysics*, [In Press]. https://arxiv.org/abs/2501.00000

Schulze-Makuch, D., Méndez, A., Fairén, A. G., von Paris, P., Turse, C., Boyer, G., … & Irwin, L. N. (2011). A Two-Tiered Approach to Assessing the Habitability of Exoplanets. *Astrobiology*, 11(10), 1041-1052. https://doi.org/10.1089/ast.2010.0592

Seager, S., Kuchner, M., Hier-Majumder, C. A., & Militzer, B. (2007). Mass-Radius Relationships for Solid Exoplanets. *The Astrophysical Journal*, 669(2), 1279-1297. https://doi.org/10.1086/521346

van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). CRC Press. https://www.taylorfrancis.com/books/mono/10.1201/9780429492259/flexible-imputation-missing-data-stef-van-buuren