IEEE ICBC'24 1570982126

ROAR: A Benchmark for NFT Rarity Meters

***Abstract*—Rarity meters are incorporated by industry and discursive by academia. Rarity, as an intuitive term, attracted numerous researchers to present their own view of it. While there is existing literature on comparing rarity meters, it requires access to NFT collection data, which can be challenging for researchers without a background in blockchain technology. This has created a demand for an easily accessible rarity meter benchmark. In this paper, we introduce the Rating over all Rarities (ROAR) benchmark, which includes data from one hundred popular NFT collections from the Ethereum blockchain, implemented a weighted correlation-based performance measurement function, as well as four state-of-the-art rarity meters (Rarity.tools, Kramer, OpenRarity, and NFTGo), along with a new rarity meter called ROAR. Our experiments show that the ROAR rarity meter, an ensemble of the other four meters, outperforms its competitors, with Rarity.tools and Kramer as runner-ups. The ROAR benchmark is a tool for examination and testing of rarity meter ideas, and we challenge readers to develop models that can outperform the ROAR rarity meter.**

***Index Terms*—Blockchain, NFT, Rarity, Benchmark, Dataset, Interpretability**

I. Introduction

Non-fungible tokens (NFTs) have become a popular form of digital asset, with each token representing a unique identifier often linked to a digital object such as a piece of artwork [1]–[3]. These NFTs can be found in various collections within a single smart contract, and are traded daily on platforms such as OpenSea, Solanart, and Singular across different blockchains including Ethereum, Solana, and Kusama [4], [5].

The value of NFTs is determined by a set of unique traits, that define each token within a collection (see Figure 1). While all NFTs are inherently unique, their trade values can vary significantly based on factors such as rarity. This has led to a growing interest in measuring and quantifying the rarity of NFTs, both within the industry and in academic research.

Rarity meters have been developed to measure the rarity of NFTs [6], but comparing these meters can be challenging without an convenient access to the underlying NFT collection data [7]. As a result, there is a need for an easily accessible benchmark for measuring rarity, which we address with the Rating over all Rarities (ROAR) benchmark. ROAR benchmark aims to provide a standardized and accessible way to evaluate the rarity of NFTs, catering to both industry professionals and researchers alike.

The main contributions of our paper to the rarity meter design problem are as follows:

1) Introduction of the Rating over all Rarities (ROAR) benchmark, which serves as a tool for early-stage examination and testing of rarity meter ideas. The ROAR benchmark will be openly available on Github once the paper is accepted.
2) Collection of trait and trade data for one hundred popular NFT collections from the Ethereum blockchain.
3) Provision of formulas and their implementations of Rarity.tools, Kramer, OpenRarity, and NFTGo state-of-the-art rarity meters in a single place.
4) Introduction of the ROAR rarity meter, which is an ensemble of the four above-mentioned state-of-the-art rarity meters, and outperforms its competitors.

The rest of the paper is organized as follows. Section II considers the related work. We introduce the formal notations and definitions used in Section III, providing details on performance measurement and state-of-the-art rarity meters. Section IV outlines the data mining methodology. ROAR benchmark is presented in Section V with a dataset and performance evaluation protocol being the main parts. The state-of-the-art rarity meters verification and ROAR benchmark showcase are in Section VI. Finally, Section VII concludes the paper and outlooks the research.

Fig. 1. NFT $n = 173$ is part of the Bored Ape Yacht Club (BAYC) collection, which consists of $10,000$ unique tokens. The contract address for this collection is `0xbc4ca0eda7647a8ab7c2061c2e118a18a936f13d`. BAYC tokens have seven distinct traits: background, clothes, eyes, hats, earrings, fur, and mouth. For NFT $n = 173$, the trait values are as follows: Yellow, Guayabera, Bored, Bayc Flipped Brim, None, Blue, and Bored

II. Related Work

Numerous research studies have been conducted on the topic of NFTs since their surge in popularity in 2021. These studies cover a range of areas, including market structure, performance, and pricing.

One notable website, Rarity.tools, has developed a formula for ranking generative art and collectible NFTs based on their

rarity [6]. This formula has gained popularity and has been adopted by various services [8].

In one of the earliest studies, researchers examined NFTs on the Ethereum and WAX blockchains to understand the market structure [9]. They collected purchase transactions and metadata from different NFT collections spanning from 2017 to 2021. The data covered six categories of NFTs: art, collectibles, games, metaverse, utility, and others. The researchers then constructed two directed graphs: one representing the network of trades and the other representing the network of NFTs. The analysis revealed that a majority of market participants specialize in a small number of preferred NFT collections, and a small fraction of traders dominate the majority of trades. Additionally, the researchers used AlexNet to encode images associated with tokens into a vector space, finding that tokens within the same collection had lower cosine distances compared to tokens from different collections. This supported the idea of visual homogeneity within most collections. The researchers also explored NFT pricing by fitting a linear regression model and found that trading history was a good predictor of future sale price, while visual properties were not.

Another study focused on NFT pricing mechanisms and the impact of Rarity.tools rarity on prices [10]. The researchers compiled a dataset of Ethereum NFT collections and their trade history from 2018 to 2022. They defined an interpretable trait-based rarity score for NFTs based on the rarity scores of their traits. The rarity score of a trait was calculated as the inverse fraction of tokens sharing that trait in the collection. The researchers also ranked NFT rarity based on the rarity rank of their rarest trait. The analysis revealed a positive correlation between rarity and sale price, as well as a negative correlation between rarity and sales volume.

While Rarity.tools has gained popularity, it is important to note that its formula lacks theoretical justification and may not be the only optimal approach. This has led to the development of other rarity meters, such as the Kramer project, which was introduced for the Kanaria NFT collection hackathon in 2021 [11]. The Kramer project proposed a workflow to compare rarity meters using trade data, a machine learning pipeline to train an optimal rarity meter, and a tournament score rarity meter. The details of the Kramer project can be found in [7]. Kramer computes a weighted correlation between pairwise relative rarities and sell prices for recent deals to compare rarity meters. The machine learning approach fits the rarity meter to the empirical data using a parametric family of rarity meters. Additionally, the tournament score is designed to meet experts' expectations of rarity using several toy examples. Although the Kramer project has been extended to two NFT collections on the Tron blockchain [12], it is still primarily an academic construction and has not been widely adopted by the industry.

OpenSea, the largest NFT marketplace by trade volume [13], in collaboration with Curio, icy.tools, and PROOF, has introduced their unique perspective on rarity named OpenRarity [14]. OpenRarity utilizes a community-driven formula that focuses on traits entropy. Subsequently, OpenRarity modified their formula to incorporate unique items and the number of non-default traits heuristics to align with expert expectations in practical examples [15]. This latest formula is now utilized by OpenSea.

From an economic standpoint, a study has been conducted to explore the investment performance of the CryptoPunks NFT collection [16], similar to the analysis done for art or real estate markets. The authors applied hedonic regression to investigate pricing, with traits and trait count being among the variables used. The results obtained supported the notion that rarity plays a significant role, as rare traits have a larger impact on price.

In addition, a new approach called NFTVis [17] has been suggested to assess the performance of NFTs based on visual features. This method estimates a token's rarity separately based on its traits and associated image. The NFTGo [18] is used as a traits-based rarity score. NFTGo is calculated using the Jaccard distance, while the image-based score is obtained as the average difference between the considered image and others in the collection.

These studies collectively reflect the ongoing research in this area. While each work assesses rarity in different ways, they all arrive at the conclusion that rarity has a substantial impact on NFT pricing.

III. Rarity Meters Design

The inherent feature of NFTs in a collection is their rarity, which can be measured using various methods. These methods are commonly referred to as rarity meters, with the rarity scores representing the values for specific NFTs. However, different rarity meters may provide different rarities, making it necessary to have a methodology for comparing them.

To address this issue, we use the methodology from [7], with slight modifications to their rarity meters notations. Our focus is on ensembles [19] of rarity meters, with the composition model referred to as an ensemble rarity meter and its building blocks as base rarity meters. We only use interpretable base rarity meters to ensure that we stay within the class of interpretable rarity meters.

While most rarity meters provide real-valued rarity scores, some may only output ranks (e.g., OpenRarity with a double sort heuristic). In such cases, we introduce consistent real-valued modifications. Additionally, even natively real-valued rarity scores may not be optimized for comparison and can be improved by tuning parameters such as scale and shift. We explore several possible extensions in later sections.

This Section introduces the formal notations and definitions used in our paper.

A. Performance Measure

Consider an NFT collection with N tokens and T traits (also known as attributes or features), denoted as $\mathbb{X}_N = \{X_n\}_{n=1}^N$, where $X_n = (x_{n,1}, \ldots, x_{n,T})$.

Definition 1. A **rarity meter** for collection $\mathbb{X}_N$ is an arbitrary function $R : \mathbb{X}_N \to [0, \infty)$, where a higher value $R(X)$ for a

given token $X \in \mathbb{X}_N$ indicates greater rarity according to the rarity meter R.

Let D denote the number of deals made with tokens from the collection $\mathbb{X}_N$ by a given time, denoted as $\{(t_d, i_d, p_d)\}_{d=1}^{D}$. Each deal is characterized by the time t_d, the index of the traded token in the collection i_d, and the deal price p_d.

The price of the token is denominated in cryptocurrency, which is subject to volatility along with fluctuations in demand. To ensure accurate comparisons, we analyze relative and pairwise prices for transactions occurring within a similar timeframe rather than relying solely on absolute and individual prices. We pair up deals in the form (j, m), where $1 \le j < m \le D$, resulting in $P = \frac{(D-1)D}{2}$ pairs.

Let

- $\varphi(R_1, R_2) = \ln(\frac{1+R_1}{1+R_2})$ be the relative rarity
- $\psi(p_1, p_2) = \ln(\frac{p_1}{p_2})$ be the relative price
- Epanechnikov kernel

$$k(t_1, t_2) = \frac{3}{4}\left(1 - \frac{|t_1 - t_2|}{h}\right)^2$$

for $|t_1 - t_2| < h = 7$ days and 0 otherwise, be weight function.

Denote $\vec{\varphi}, \vec{\psi}, \vec{k} \in \mathbb{R}^{P\times 1}$ vectors of all relative rarities, prices and weights, i.e. the pth components corresponding to the deals pair (j, m) equals

$$\begin{aligned}\varphi_p &= \varphi(R(X_{i_j}), R(X_{i_m}))\\ \psi_p &= \psi(p_j, p_m)\\ k_p &= k(t_j, t_m).\end{aligned}$$

Denote the weighted correlation function [20] of $\vec{\varphi}$ and $\vec{\psi}$ with weights $\vec{k}$ as $\mathrm{corr}(\vec{\varphi}, \vec{\psi}; \vec{k})$. Specifically, we denote $\vec{\mathbf{1}}_J \in \mathbb{R}^{J\times 1}$ as the vector of all ones, and for an arbitrary vector $\vec{a} \in \mathbb{R}^{J\times 1}$, we define the mean vector with weights $\vec{k}$ as $\overline{\vec{a}}_{\vec{k}} = \frac{\vec{k}^T\vec{a}}{\vec{k}^T\vec{\mathbf{1}}_J} \cdot \vec{\mathbf{1}}_J$, and the deviation vector as $\dot{\vec{a}}_{\vec{k}} = \vec{a} - \overline{\vec{a}}_{\vec{k}}$. Overall, correlation function is written as follows:

$$\mathrm{corr}(\vec{\varphi}, \vec{\psi}; \vec{k}) = \frac{\dot{\vec{\varphi}}_{\vec{k}}^T \cdot \mathbf{K} \cdot \dot{\vec{\psi}}_{\vec{k}}}{\sqrt{\dot{\vec{\varphi}}_{\vec{k}}^T \cdot \mathbf{K} \cdot \dot{\vec{\varphi}}_{\vec{k}}} \cdot \sqrt{\dot{\vec{\psi}}_{\vec{k}}^T \cdot \mathbf{K} \cdot \dot{\vec{\psi}}_{\vec{k}}}}.$$

Definition 2. The **performance measure** F of the rarity meter R on the collection $\mathbb{X}_N$ is defined as

$$F(R; \mathbb{X}_N) = \mathrm{corr}(\vec{\varphi}, \vec{\psi}; \vec{k}). \tag{1}$$

The performance measure F falls within the range of $[-1, 1]$, where a higher value of $F(R; \mathbb{X}_N)$ indicates better performance of the rarity meter R on the collection $\mathbb{X}_N$.

B. State-of-the-Art

Let us consider the existing approaches to rarity meters. All rarity meters are formulated for the collection $\mathbb{X}_N = \{X_n\}_{n=1}^{N}$, where $X_n = (x_{n,1}, \ldots, x_{n,T})$. If the NFT X_n has no value for the trait t, we set it as an artificial `None` value.

Denote the multiset of tth attributes in the collection $\mathbb{X}_N$ as $\mathbb{X}_{N,t} = \{x_{n,t}\}_{n=1}^{N}$. Also, denote the cardinality of a set A as $\#A$.

In addition to attributes defined by the developers of the NFT collection, rarity meters can add artificial attributes to factor into ranking. We will refer to such synthetic traits as **meta-traits**.

1) Rarity.tools: The approach computes individual scores for all traits and provides the resulting rarity as their sum. The individual scores are inverse fractions of the given trait rarity value in the collection. In addition, Rarity.tools preprocesses the collection by adding a $(T+1)$st meta-trait called **traits count**, representing the number of not-`None` traits of an individual token. Let us denote $[A]$ the indicator of the event A, where $[A] = 1$ if A is true, and $[A] = 0$ if A is false. The **traits count** $x_{n,T+1}$ is computed as follows

$$x_{n,T+1} = \sum_{t=1}^{T} [x_{n,t} \neq \texttt{None}]. \tag{2}$$

Definition 3. The **Rarity.tools rarity meter** R_{rt} for a given token $X_k \in \mathbb{X}_N$ equals

$$R_{rt}(X_k) = \sum_{t=1}^{T+1} \texttt{score\_rt}_t(X_k),$$

where for $t = 1, \ldots, T+1$:

$$\texttt{score\_rt}_t(X_k) = \frac{N}{\#\{x \in \mathbb{X}_{N,t} | x = x_{k,t}\}}.$$

2) Kramer: The approach computes individual scores for all traits and provides the resulting rarity as their weighted sum with coefficients that maximize the performance measure F. The individual scores compare between groups of traits in a tournament style, where the fewer tokens have a given trait value, the higher the score this group receives. Let the number of unique tth trait values be denoted as G, with $N_1, \ldots, N_G$ being the numbers of entries such that $\sum_{g=1}^{G} N_g = N$. In the pairwise comparison of groups i and j, where $1 \le i < j \le G$, group i gains $\frac{N_j}{N_i+N_j}$ points. As a result, the group k of size N_k gains

$$a_k = \frac{1}{G-1} \cdot \left(-\frac{1}{2} + \sum_{g=1}^{G} \frac{N_g}{N_k + N_g}\right)$$

points, where $-\frac{1}{2}$ is needed to eliminate the tournament within the same group as $\frac{N_k}{N_k+N_k} = \frac{1}{2}$ from the sum.

After calculating all group averages, we divide them by the attribute average calculated for all values of this attribute:

$$s_k = \frac{a_k}{\frac{1}{N} \cdot \sum_{i=1}^{G} N_i \cdot a_i}.$$

The individual score of the token X_n for the trait t is $\texttt{score\_kr}_t(X_n) = s_k$, where k is the index of the $x_{n,t}$'s group.

Let the set of $\texttt{score_kr}_t$s' non-negative combinations be denoted as $\mathcal{R}$:

$$\mathcal{R} = \{R | R = \sum_{t=1}^{T} \alpha_t \cdot \texttt{score\_kr}_t, \alpha_t \geq 0, t = 1, \ldots, T\}.$$

Definition 4. The **Kramer rarity meter** R_{kr} for a given token $X_k \in \mathbb{X}_N$ equals

$$R_{kr}(X_k) = \sum_{t=1}^{T} \alpha_t \cdot \texttt{score\_kr}_t(X_k),$$

where $\alpha_1, \ldots, \alpha_N$ are the solution of

$$F(R; \mathbb{X}_N) \to \max_{R \in \mathcal{R}}.$$

3) OpenRarity: The approach computes the Shannon information (also known as information content) of a token and normalizes it by the collection average. The authors do not provide a formula, but instead offer several extra heuristics. Therefore, we provide our interpretation of the OpenRarity meter and further justify it numerically.

Firstly, OpenRarity ignores `None` traits. To support this idea in formulas, we will use indicator notations. For $X_n \in \mathbb{X}_N$ and $t = 1, \ldots, T$:

$$P(x_{n,t}) = \frac{\#\{x \in \mathbb{X}_{N,t} | x = x_{n,t}\}}{N}$$

$$I(X_n) = \sum_{t=1}^{T} -\log\left(P(x_{n,t})\right) \cdot [x_{n,t} \neq \texttt{None}]$$

$$\mathbb{E}I(X) = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} -\log\left(P(x_{n,t})\right)) \cdot [x_{n,t} \neq \texttt{None}]$$

The original OpenRarity claimed to compute the resulting rarity as

$$R_{or,0}(X_n) = \frac{I(X_n)}{\mathbb{E}I(X)}.$$

However, they introduced two heuristics:

1) **Double Sort:** introduce the meta-trait of the number of unique traits

$$x_{n,0} = \sum_{t=1}^{T} [\#x \in \mathbb{X}_{N,t} | x = x_{n,t} = 1]$$

and sort by the pair meta-trait and Shannon information, where Shannon information only matters when comparing two tokens with an equal number of unique traits.
2) **Trait Count:** introduce the meta-trait of the number of not-`None` traits $x_{n,T+1}$ (2) and compute OpenRarity on the top of $T+1$ traits $t = 1, \ldots, T+1$.

The double sort heuristic makes ranking instead of metering, so we modify the formula with the additional term to order by unique traits. As the maximum Shannon information is the NFT collection setting equals $I_{\max} = (T+1)\log(N)$ and it only achieves this for NFT's with the unique traits, the additional term is

$$\texttt{reg}(X_n) = x_{n,0} \cdot \frac{(T+1)\log N}{\mathbb{E}I(X)}.$$

Definition 5. The **OpenRarity rarity meter** R_{or} for a given token $X_k \in \mathbb{X}_N$ is defined as:

$$\begin{aligned} R_{or}(X_k) &= R_{or,0}(X_k) + \texttt{reg}(X_k) \\ &= \frac{I(X_k) + x_{k,0} \cdot (T+1)\log N}{\mathbb{E}I(X)}. \end{aligned}$$

4) NFTGo: The approach calculates the Jaccard distance in the trait space between a specific NFT and the entire collection. The Jaccard distance is determined by the ratio of the set intersection and set union.

Let the Jaccard distance between two tokens X_i and X_j in $\mathbb{X}_N$ be denoted as $\texttt{JD}(X_i, X_j)$. It can be calculated using the formula:

$$\texttt{JD}(X_i, X_j) = \frac{\sum_{t=1}^{T} [x_{i,t} = x_{j,t}]}{2T - \sum_{t=1}^{T} [x_{i,t} = x_{j,t}]}.$$

It follows that for any pair X_i and X_j in $\mathbb{X}_N$, $\texttt{JD}(X_i, X_j) \in [0, 1]$.

Definition 6. The **NFTGo rarity meter** R_{go} for a given token $X_k \in \mathbb{X}_N$ equals

$$R_{go}(X_k) = \text{normalize}\left(\sum_{n=1}^{N} (1 - \texttt{JD}(X_k, X_n))\right),$$

where normalize maps R_{go} in such a way that the minimum value on the collection is 0 and the maximum value equals 100.

IV. Data Mining

The benchmark for NFT collections is determined by the presence of traits and the execution of trades. Traits are established during collection creation or individual NFT minting, while trades occur through auctions, bids, or direct deals over time. Typically, collection creators and marketplaces set a fee ranging from 0 to 10 percent.

Platforms such as OpenSea, Solanart, and Singular facilitate the daily trading of thousands of NFT collections, utilizing blockchains like Ethereum, Solana, and Kusama [4], [5]. Collections are smart contract instances with openly available code and transaction history, enabling the collection of necessary data and its interpretation based on smart contract-defined rules. Different blockchains have distinct application programming interfaces (APIs) and smart contract languages, resulting in variations in smart contract structure even within the same language.

In order to ensure a representative sample of collections, we focused on one hundred collections, with an emphasis on Ethereum due to its superior performance in terms of capitalization, number of deals, owners, and average price [21], and to streamline the data mining process. Ethereum smart contracts are written in the Solidity programming language, with various optional standards [2], [22]–[24]. The Ethereum Improvement Proposal (ERC) 721 covers NFT needs and is the most popular among collections. We specifically excluded collections with ERC 1155 standard smart contracts when they do not have traits and supply in metadata.

While anyone can launch their own Ethereum node to access the required data, this process is resource-intensive and redundant for our purposes as we only need a small portion of data related to a hundred smart contracts with known addresses. The QuickNode API [25] provided metadata parsed into traits and general collection information. Working with metadata involved a semi-manual consideration of each collection. To ensure data consistency, we cross-checked the collection size (total and actual supply) and traits with OpenSea and Etherscan. Additionally, Alchemy [26] and Covalent [27] provided trade data in parallel. Collecting trade data was less challenging as both sources were consistent and aligned with the data on Etherscan.

NFTs are commonly associated with visual arts, however, in our research, images were not a focus and were not collected.

V. ROAR Benchmark

To standardize the comparison of different rarity meters, we introduce the Rating Over All Rarities (ROAR) benchmark. Upon acceptance, the benchmark and its Python code will be openly available on Github once the paper is accepted. The benchmark comprises of a dataset and a performance evaluation parts (see figure 2), which are discussed in this Section.

A. Dataset

The current dataset consists of 100 NFT collections. The list of these collections can be found in the file at the root of the project. Each collection is identified by its name, short name (symbol), smart contract instance address, and total supply size N. The symbols are used to distinguish between collections in the benchmark.

Each collection contains traits and trade data, which are stored in corresponding subfolders in the dataset folder. Each collection has its own file in both subfolders. The traits table contains the NFT identifiers followed by the trait names in the first row. The remaining lines represent each NFT, with one line per NFT. The trades file contains one line per transaction, with information such as transaction hash, block height, NFT identifier, and price in Ether (ETH). While the transaction hash is not used for performance evaluation, it is useful for ensuring data consistency. The block height is used to determine the time t of the transaction, as transactions are committed in blocks and each block has a timestamp. The block timestamps file can be found in the utils folder.

The collections in the dataset contain anywhere from 2027 to 28 170 NFTs and from 2 to 67 traits. The traits are immutable for all collections, while trades occur over time. The entire transaction history was covered up to November 2023. The trade data contains 971 to 130,426 trades per collection. The root folder also contains a file with the latest parsed block, which is the same for all collections.

B. Performance Evaluation

Some rarity meters require training. In order to compare all methods on an equal footing, we divide the trade data for each collection into two parts based on time. Seventy percent of the earliest transactions are used for training, while thirty percent of the later transactions are used for testing. Since each collection has a different creation time and transaction activity pattern, the train and test time frames and durations vary.

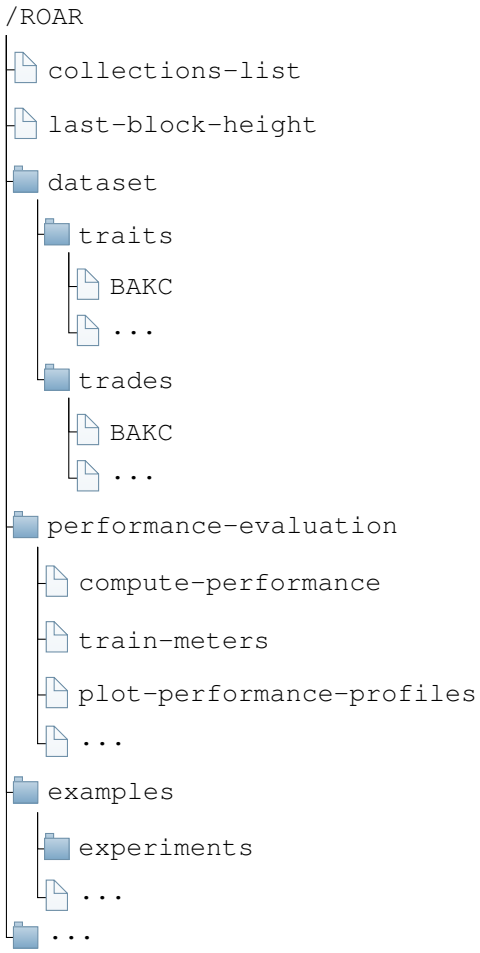

Fig. 2. Benchmark repository directory tree

For each collection, the rarity meter is trained using trait and trade data, and its performance is evaluated on the test trade data. The benchmark calculates one performance measurement for each pair of rarity meter and NFT collection.

To visualize the comparison of rarity meters on the entire dataset, we use performance profiles [28]. These profiles can accommodate different performance measures and distances, but we specifically formulate them for the performance measure F (1) and use the absolute difference as the distance. Let there be M rarity meters $R_1, \ldots, R_M$ and C NFT collections $\mathbb{X}^1, \ldots, \mathbb{X}^C$. The performance of the mth rarity meter R_m on the cth collection is denoted as $F_{m,c} = F(R_m; \mathbb{X}^c)$.

Definition 7. The **performance profile** $\rho\colon [0,2] \to [0,1]$ of a rarity meter R_m in a rarity meters set $R_1, \ldots, R_M$ on ROAR

benchmark is defined as:

$$\rho(\tau; R_m) = \frac{1}{C} \sum_{c=1}^{C} \left[F_{m,c} + \tau \geq \max\{F_{i,c}\}_{i=1}^{M} \right],$$

where $[A]$ is the indicator of an event A.

The performance profile $\rho(\tau; R_M)$ shows the fraction of collections where the performance measure F for rarity meter R_m is no more than τ away from the best performing rarity meter on that collection. Since $\rho(\tau; R_M)$ is a fraction, its output range is $[0, 1]$. The performance measure F is a weighted correlation and is bounded between -1 and 1, so the best possible performance is 1, and all runner-ups have a performance no more than 2. Therefore, the input range for ρ is $[0, 2]$. Additionally, since the indicator

$$\left[F_{m,c} + \tau \geq \max\{F_{i,c}\}_{i=1}^{M} \right]$$

is monotonic with τ, the performance profile ρ is also monotonic with τ. A higher value for ρ indicates a better performance for the rarity meter, with the point $(0, 1)$ indicating the best performance for all collections.

Inspired by state-of-the-art (Section III-B), we propose our own rarity meter as an ensemble of other rarity building blocks. Let $\mathcal{R}_{rr}$ be the set of non-negative combinations of $\texttt{score\_rt}_t$, $\texttt{score\_kr}_t$, $\texttt{score\_or}_t = -\log\left(P(x_{k,t})\right) \cdot [x_{k,t} \neq \texttt{None}]$, $\texttt{traits\_coint}$, and R_{go}.

Definition 8. The **ROAR rarity meter** R_{rr} for a given token $X_k \in \mathbb{X}_N$ equals

$$R_{rr}(X_k) = \arg\max_{R \in \mathcal{R}_{rr}} F(R; \mathbb{X}_N).$$

Currently, ROAR benchmark contains five rarity meters: Rarity.tools, Kramer, OpenRarity, NFTGo (definitions 3-6) and ROAR. The benchmark computes performance measurements matrix $(F_{m,c})_{m,c=1}^{M,C}$ on a test data and plots performance profiles. More examples of dataset usage and performance evaluation can be found in the examples folder.

VI. Numerical Experiments

To demonstrate the practicality of our benchmark, we conducted a quantitative analysis. The code used to reproduce these experiments can be found in the experiments subfolder of the examples folder (see figure 2).

A. State-of-the-Art Verification

As part of our benchmark showcase, we also performed a set of verification experiments to ensure the accuracy of our implementation and interpretation of the state-of-the-art methods (see Section III-B).

1) Rarity.tools: The web site provides graphical user interface (GUI) with NFT's rarity score and each traits' impact on it. We selected two collections at random from our dataset and analyzed five tokens from each collection. Our results were consistent with the rarity scores and trait impacts reported on the website, as well as with the formula described in [6].

2) Kramer: The detailed description of the meter is provided in [7] and the code with the computation results is given in [11] and [12]. The scores are provided for the Kanaria collection on Kusama blockchain, and BAYC and MAYC collections on TRON blockchain. We omitted the Kanaria collection due to its collection-specific design elements, such as edition and sets. We verified the rarity scores for BAYC and MAYC collections on TRON and found that the trait weights were consistent up to the fourth decimal place, which is acceptable given the numerical optimization involved.

We also computed rarity scores for BAYC and MAYC collections on Ethereum, which have the same traits but different trades. The resulting trait weights were very similar, with less than one percent deviation. This suggests that both the Kramer rarity meter and user preferences for rarity perception are robust.

3) OpenRarity: The OpenRarity formula is introduced in [14] and is followed by two heuristics [15]: double sort and trait count. The double sort heuristic does not support rarity scores natively but only rarity ranks. As a result, the OpenSea GUI and API provide ranks only.

We modified the initial OpenRarity formula to have scores resulting in (3). Afterward, we selected two collections at random, computed rarity scores and ranks, collected ranks by OpenSea, and convinced that the results are equal.

4) NFTGo: The website offers both a graphical user interface (GUI) and an application programming interface (API) for determining the rarity score of NFTs. We utilized API platform to compare the results obtained. In a small percentage of collections (less than five percent) and an even smaller percentage of tokens (less than one percent), we observed a deviation in the scores. This discrepancy suggests that either the method used for production has been modified or there are errors in the trait collection process. However, we are unable to determine the exact cause at this time.

B. Rarity Meters Benchmark

The performance profiles (definition 7) of the four state-of-the-art methods Rarity.tools, Kramer, OpenRarity, NFTGo (Section III-B), and the new ROAR (definition 8) ensemble method are shown in figure 3. ROAR outperforms the other methods, being the best in more than 50% of collections and never losing by more than 0.2. Kramer and Rarity.tools come in second and third place, respectively.

We also computed performance profiles for the training dataset (see figure 4). The results are even better for ROAR and Kramer as they use training trades to tune their parameters. Since the ROAR's search space $\mathbb{R}_{rr}$ includes all the other rarity meters, it provides the best training performance for all the collections. While the four state-of-the-art methods give the best solutions for 20% of collections, ROAR rarity meter gives the unique best solution for 80% of collections or more.

C. OpenRarity Heuristics Quantification

Numerical analysis and examples play a crucial role in introducing the heuristics for OpenRarity [15]. However, no

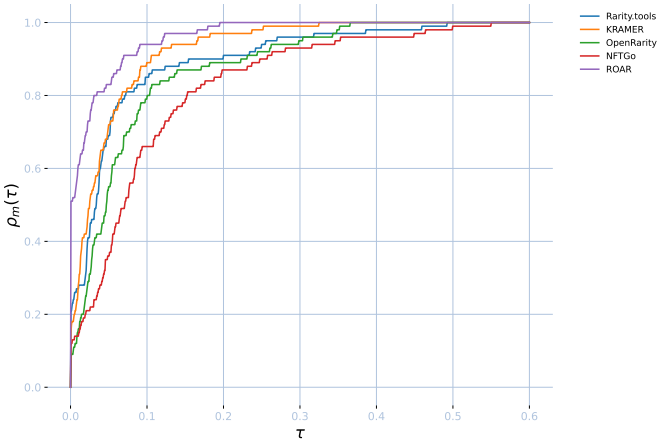

Fig. 3. Rarity meters performance profiles for ROAR benchmark

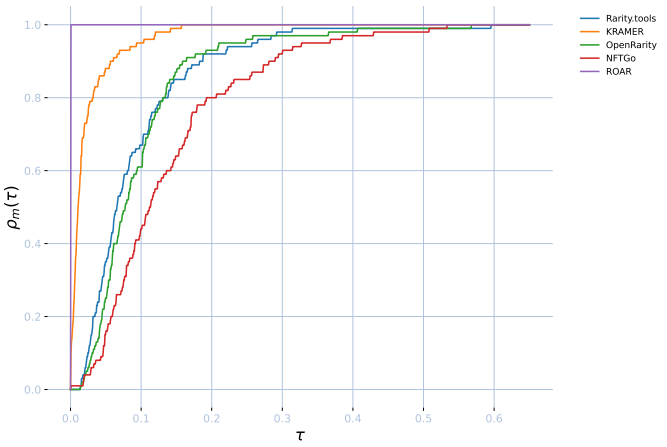

Fig. 4. Rarity meters performance profiles for ROAR benchmark's train data

specific quantitative criteria were provided for decision making. The ROAR benchmark allows for a thorough examination of these heuristics. Figure 5 presents the performance of the `original` OpenRarity, as well as modifications using the `double sort` and `trait count` heuristics, and the current `production` version which incorporates both modifications. The results show all the methods work similarly, but both heuristics improve the overall performance, with the combination of the two being the most effective. This experiment provides numerical evidence for the effectiveness of these modifications.

VII. Conclusions and Future Work

In 2023, rarity meters have become an essential component of the NFT industry, allowing people to assess the rarity of NFTs despite their fluctuating prices. This paper introduces a benchmark called Rating over all rarities (ROAR), which builds upon the comparison methodology from [7] by incorporating one hundred NFT collections from Ethereum and performance profiles [28] to visualize multiple measurements. ROAR serves as a comparison tool for designers and researchers to evaluate the performance of different rarity meters, including Rarity.tools, Kramer, OpenRarity, and NFTGo.

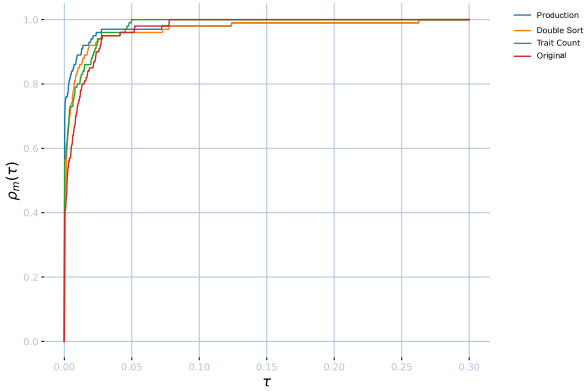

Fig. 5. Performance profiles of OpenRarity rarity meters modifications for ROAR benchmark

The ROAR benchmark not only transforms the design of rarity meters from an art to an optimization task but also facilitates the examination and testing of new ideas at an early stage. For example, in Section VI-C, the benchmark is used to numerically evaluate the impact of modifications to OpenRarity.

Additionally, a new rarity meter called ROAR rarity meter is introduced, which is a linear ensemble of the state-of-the-art rarity meters and outperforms the others. This showcases the potential for further advancements in this field through mathematical and computer science innovations, challenging researchers to surpass this model.

It is important to note that both state-of-the-art rarity meters and our model are limited to interpretable rarity meters, where the resulting rarity can be explained through combinations of initial traits. Without a performance measure, researchers must justify the meter design. However, as shown in Section VI-C, a performance measure makes the explanation optional. If we relax the interpretability constraint and allow for non-interpretable rarity meters, the search space becomes larger and the results are expected to be even better. Additionally, standard machine learning techniques such as neural networks and Gaussian processes can be used to find effective non-interpretable rarity meters.

Within interpretable models, various coefficients are of interest. For example, relative rarity φ in (1) varies with rarity scale. Thus, training the scale for Rarity.tools, OpenRarity, and NFTGo can improve their results. Additionally, as coefficients are trained on the training data, they may differ for different training sets. This highlights the importance of training robustness and motivates regularization for Kramer and ROAR rarity meters to provide stable and sparse solutions.

Some deals may have a significant negative impact on performance. This not only indicates a poorly designed rarity meter, but also an unusual deal–an outlier. Such outliers may be related to suspicious activity, making rarity meters with performance evaluation a potential alternative to transfer graph

pattern [29] and community detection [30] for detecting price pumps.

Real-valued rarity scores can be useful for some users and for evaluating performance, but they may be frustrating for others. The performance measure (2) that has been implemented is subjective, as it changes with the scaling of rarity meters. This is due to the non-negative output space of rarity meters, which results in a constant regularization term in the relative rarity φ. Rarity ranks offer less detailed information to users, but their class is larger than rarity scores, potentially containing unique and interesting meters. Exploring alternative performance measures for both scores and ranks is a potential area for future research, with Spearman's rank correlation coefficient serving as a potential starting point for rank-based performance measurement.

REFERENCES

[1] L. Oliveira, I. Bauer, L. Zavolokina, and G. Schwabe, "To token or not to token: Tools for understanding blockchain tokens," in *International Conference on Information Systems 2018, ICIS 2018*, 2018, pp. 1–17.

[2] M. di Angelo and G. Salzer, "Tokens, Types, and Standards: Identification and Utilization in Ethereum," in *2020 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPS)*. IEEE, 8 2020, pp. 1–10. [Online]. Available: https://ieeexplore.ieee.org/document/9126009/

[3] N. Wang, S. Chi-Kin Chau, and Y. Zhou, "Privacy-Preserving Energy Storage Sharing with Blockchain," in *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*. New York, NY, USA: ACM, 2021. [Online]. Available: https://doi.org/10.1145/3447555.3464869

[4] G. Wood, "Polkadot: Vision for a Heterogeneous Multi-Chain Framework," *Whitepaper*, pp. 1–21, 2017. [Online]. Available: https://github.com/w3f/polkadot-white-paper/raw/master/PolkaDotPaper.pdf

[5] A. Yakovenko, "Solana: A new architecture for a high performance blockchain," p. 32, 2018. [Online]. Available: https://solana.com/solana-whitepaper.pdf

[6] Rarity.tools, "Ranking Rarity: Understanding Rarity Calculation Methods," 2021. [Online]. Available: https://raritytools.medium.com/ranking-rarity-understanding-rarity-calculation-methods-86ceaeb9b98c

[7] M. Krasnoselskii, Y. Madhwal, and Y. Yanovich, "KRAMER: Interpretable Rarity Meter for Crypto Collectibles," *IEEE Access*, vol. 11, pp. 4283–4290, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10014994/

[8] Slashdot Media, "Best rarity.tools Alternatives & Competitors," 2022. [Online]. Available: https://sourceforge.net/software/product/rarity.tools/alternatives

[9] M. Nadini, L. Alessandretti, F. Di Giacinto, M. Martino, L. M. Aiello, and A. Baronchelli, "Mapping the NFT revolution: market trends, trade networks, and visual features," *Scientific Reports*, vol. 11, no. 1, p. 20902, 12 2021.

[10] A. Mekacher, A. Bracci, M. Nadini, M. Martino, L. Alessandretti, L. M. Aiello, and A. Baronchelli, "Heterogeneous rarity patterns drive price dynamics in NFT collections," *Scientific Reports*, vol. 12, no. 1, p. 13890, 8 2022. [Online]. Available: https://www.nature.com/articles/s41598-022-17922-5

[11] M. Krasnoselskii, Y. Madhwal, and Y. Yanovich, "KRAMER: Kanaria NFT Collection Rarity Meter," in *2022 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*. IEEE, 5 2022, pp. 1–2. [Online]. Available: https://ieeexplore.ieee.org/document/9805542/

[12] M. Krasnoselskii, Y. Madhwal, A. Stepin, and Y. Yanovich, "NFT SMASH: Game to Test Your NFT Rarity Sense," in *2023 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*. IEEE, 5 2023, pp. 1–2.

[13] B. White, A. Mahanti, and K. Passi, "Characterizing the OpenSea NFT Marketplace," in *Companion Proceedings of the Web Conference 2022*. New York, NY, USA: ACM, 4 2022, pp. 488–496.

[14] OpenRarity, "Introducing OpenRarity," 2022. [Online]. Available: https://mirror.xyz/openrarity.eth/-R8ZA5KCMgqtsueySlruAhB77YBX6fSnS_dT-8clZPQ

[15] ——, "Proposal: Add Trait Count to OpenRarity," 2022. [Online]. Available: https://mirror.xyz/openrarity.eth/oNo7AmgXopMCKq95gv_Xe0p5pKQ_qHWKzTV11DpFxgE

[16] L. Schaar and S. Kampakis, "Non-fungible Tokens as an Alternative Investment: Evidence from CryptoPunks," *The Journal of The British Blockchain Association*, vol. 5, no. 1, pp. 1–12, 2022.

[17] F. Yan, X. Wang, K. Mao, W. Zhang, and W. Chen, "NFTVis: Visual Analysis of NFT Performance," in *2023 IEEE 16th Pacific Visualization Symposium (PacificVis)*. IEEE, 4 2023, pp. 82–91.

[18] NFTGo, "NFTGo: The World's Leading NFT Data Intelligence Provider." [Online]. Available: https://docs.nftgo.io/

[19] O. Sagi and L. Rokach, "Ensemble learning: A survey," *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4, 2018.

[20] J. F. P. d. Costa, "Weighted Correlation," in *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1653–1655.

[21] DappRadar UAB, "Top NFT Collections," 2023. [Online]. Available: https://dappradar.com/rankings/nft/collections

[22] Q. Wang, R. Li, Q. Wang, and S. Chen, "Non-Fungible Token (NFT): Overview, Evaluation, Opportunities and Challenges," 5 2021. [Online]. Available: https://arxiv.org/abs/2105.07447v3

[23] V. Davydov, A. Gazaryan, Y. Madhwal, and Y. Yanovich, "Token Standard for Heterogeneous Assets Digitization into Commodity," in *Proceedings of the 2019 2nd International Conference on Blockchain Technology and Applications*. New York, NY, USA: ACM, 12 2019, pp. 43–47. [Online]. Available: https://dl.acm.org/doi/10.1145/3376044.3376053

[24] H. Benedetti and G. Rodríguez-Garnica, "Tokenized Assets and Securities," *The Emerald Handbook on Cryptoassets: Investment Opportunities and Challenges*, pp. 107–121, 1 2023.

[25] QuickNode, "API Documentation for Web3," 2023. [Online]. Available: https://www.quicknode.com/docs/welcome

[26] Alchemy, "Token API." [Online]. Available: https://www.alchemy.com/token-api

[27] Covalent, "Token Balances API." [Online]. Available: https://www.covalenthq.com/products/token-balances-api/

[28] E. D. Dolan and J. J. Moré, "Benchmarking optimization software with performance profiles," *Mathematical Programming*, vol. 91, no. 2, pp. 201–213, 1 2002. [Online]. Available: http://link.springer.com/10.1007/s101070100263

[29] M. L. Morgia, A. Mei, A. M. Mongardini, and E. N. Nemmi, "A Game of NFTs: Characterizing NFT Wash Trading in the Ethereum Blockchain," in *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 7 2023, pp. 13–24.

[30] N. Tahmasbi, G. Shan, and A. M. French, "Identifying Washtrading Cases in NFT Sales Networks," *IEEE Transactions on Computational Social Systems*, pp. 1–12, 2023.