Causal evidence for the primordiality of colours in trans-Neptunian objects

Benjamin L. Davis [1,†⋆] Mohamad Ali-Dib [1,†] Yujia Zheng [2,†] Zehao Jin [1,3,†] Kun Zhang [2,4]
and Andrea Valerio Macciò [1]

[1] *Center for Astrophysics and Space Science (CASS), New York University Abu Dhabi, PO Box 129188, Abu Dhabi, UAE*
[2] *Carnegie Mellon University, Pittsburgh, PA, USA*
[3] *Center for Astronomy and Astrophysics and Department of Physics, Fudan University, Shanghai 200438, People's Republic of China*
[4] *Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE*
[†] *These authors contributed equally to this work and are listed alphabetically.*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

The origins of the colours of Trans-Neptunian Objects (TNOs) represent a crucial unresolved question, central to understanding the history of our Solar System. Recent observational surveys revealed correlations between the eccentricity and inclination of TNOs, and their colours. This rekindled the long-standing debate on whether these colours reflect the conditions of TNO formation or their subsequent evolution. We address this question using a model-agnostic, data-driven approach that unanimously converges to a common causal graph from the analysis of two different datasets, each from two different conditional independence test methods. For evaluation, we demonstrate how our model is consistent with the currently-accepted paradigms of TNOs' dynamical histories, without involving any orbital modelling or physics-based assumptions. Our causal model (with no knowledge of the existence of Neptune) predicts the need for an unknown confounding variable, consistent with Neptune's effects. The model predicts that the colour of TNOs is the root cause of their inclination distribution, rather than the other way around. This strongly suggests that the colours of TNOs reflect an underlying dynamical property, most likely their formation location. Our model excludes formation scenarios that invoke substantial colour modification by subsequent evolution. We conclude that the colours of TNOs are predominantly primordial.

Key words: comets: Kuiper belt: general – Kuiper belt objects: asteroids: general – methods: statistical

1 INTRODUCTION

Trans-Neptunian Objects (TNOs) are invaluable probes into the history and evolution of our Solar System (Morbidelli & Nesvorný 2020). However, the wealth of information they encode is difficult to decipher with numerous, often opposing interpretations of the their properties. This includes intrinsic characteristics such as their sizes and correlated properties such as their orbits and surface photometric colours. The last two have long been closely examined in an effort to unravel the relation between them (Jewitt & Luu 2001; Liu & Ip 2019; Chen et al. 2022; Bernardinelli et al. 2025).

There are three leading theories regarding the origin of TNO colours:

(i) The primordial origin hypothesis of the TNO colour diversity argues that TNO colours reflect compositional gradients in the protoplanetary disk, preserved since formation (Brown 2012; Nesvorný et al. 2020; Ali-Dib et al. 2021; Buchanan et al. 2022; Pinilla-Alonso et al. 2025). Objects formed at different heliocentric distances thus acquired distinct volatile and refractory compositions, leading to colour variations. For example, objects that formed beyond the CO and N_2 snowlines could have acquired redder surfaces. Dynamical processes (e.g., planetary migration and scattering) later redistributed these bodies into their current orbits, imprinting correlations between colour and orbital parameters like inclination. In this scenario, using causality theory jargon, inclination (inc) is said to be caused by the colours, which is indicative of the formation location. In reality, an additional parameter, a_{ini} (the formation location), sets the inclination and colour simultaneously. The initial semimajor axis thus implies something directly about colour, and the scattering of TNOs by Neptune causes a_{ini} to be related to the present e and inc.

(ii) However, alternatively, many works (Luu & Jewitt 1996; Stern 2002; Hainaut & Delsanti 2002; Thébault & Doressoundiram 2003; Ayala-Loera et al. 2018) argued that collisional evolution is the origin of TNO colours, where collisions expose fresh subsurface ices or organic materials, altering albedo and spectral slopes. Dynamically excited populations (higher e and inc) experience more frequent collisions due to orbital crossings, leading to colour–inclination correlations. This framework treats colour as a secondary property shaped by post-formation bombardment. Opponents of this model argue that if collisional resurfacing were causal, dynamically excited populations would exhibit homogenised colours over time due to frequent mixing.

(iii) A third possibility proposed that initially diverse bulk compositions undergo selective volatile evaporation post-formation, establishing steep compositional gradients across the primordial disk that, coupled with subsequent UV photolysis and particle irradiation,

⋆ E-mail: ben.davis@nyu.edu (BLD)

yield distinct surface chemistries. A key difference between this and the ‘primordial origin’ hypothesis is the necessity of post-formation irradiation, either pre-instability (Brown et al. 2011; Wong & Brown 2016, 2017) or post-instability (Kaňuchová et al. 2012). From a causality lens, a post-instability irradiation model introduces a causal relationship between the current semimajor axis and the colour of TNOs.

All of the three theories explain the observed correlation between colour and inclination, therefore the origin of TNO colours remains a long-standing debate. However, the three models do not share the same causation: the first primordial origin model implies that colour causes the current inclination; the second collisional model, in contrary, demands inclination to cause colour; while the third post-formation theory requires not only colour to cause inclination, but also the current semimajor axis to cause colour. If one can find the causal structure between TNO colours and their orbital parameters, the three theories will be distinguishable.

Identifying cause-effect relationships is crucial for moving beyond mere correlation to uncover the underlying causal mechanisms governing a system. Traditionally, causal relationships are established through interventions or randomised experiments, where one variable is explicitly manipulated while all others are held constant, and the resulting effects are observed. However, such interventions are infeasible in fields like astronomy, where the ‘test subjects’ exist at unreachable astronomical distances. Consequently, advanced methods are required to infer causal relationships from purely observational data—an endeavour that lies at the core of causal discovery (Spirtes et al. 2001).

For decades, causal discovery has been a transformative tool in science, enabling researchers to look beyond correlation and uncover the fundamental mechanisms driving complex systems. Its applications in biology are extensive, from mapping intricate protein signalling pathways (Friedman 2004) to deciphering gene regulatory networks (Sachs et al. 2005). The methodology’s utility extends into physics, where it is invaluable for analysing systems that defy direct experimentation; examples include identifying the drivers of plasma instabilities in fusion reactors and modelling emergent causal structures within condensed matter (Runge et al. 2019). Although a newer frontier for astrophysics, recent works are beginning to demonstrate the power of causal discovery in decoding astronomical data (Pasquato et al. 2023; Pasquato 2024; Jin et al. 2024, 2025a,b; Davis et al. 2025). From cellular processes to cosmic structures, this approach provides a robust framework for modelling the underlying causal architecture of the natural world, built upon the foundational contributions of Spirtes et al. (2001) and Pearl (2009).

In this letter, we use a purely data-driven, model-agnostic, statistical causal discovery method to study the causal structure among the dynamical parameters and colours of TNOs, revealing a causal structure consistent with model one, the primordial origin model, while ruling out the other two. We show that not only this technique allows us to derive some of the main lines of the current consensus on the origins of TNOs, but also that it elucidates the direction of causality between the dynamical parameters and colours of TNOs, and predicts the existence of an unknown perturbing body, i.e., Neptune. We first give an overview of our data sample (§2), detail our causal discovery methods (§3), present the results of our analysis (§4), and finally conclude with an overall summary (§5).

2 DATA

We use two separate datasets for this work: the Col-OSSOS survey, and Dark Energy Survey (DES). Both include hot classicals, centaurs, and resonant/scattered objects. For each TNO (see Fig. 1), we have three orbital elements: semimajor axis (a), eccentricity (e), and inclination (inc); and we have spectral slope (i.e., colour). A fundamental assumption of this work is that colours are primordial, and thus strongly correlated to the initial location of a TNO. Hereafter, we treat colours as a proxy for the initial semimajor axis of the objects. In the following, the two datasets are analysed separately as they originate from different surveys with different characteristics and observational biases.

2.1 The Colours of the Outer Solar System Origins Survey

Our first dataset is based on (but not exclusively) the Colours of the Outer Solar System Origins Survey (Col-OSSOS; Schwamb et al. 2019). It was originally assembled by Marsset et al. (2019) and then re-used by Ali-Dib et al. (2021). It contains a total of 229 TNOs in a dataset for which discovery biases were modelled. These consist of Hot Classicals (48), Resonant (102), Centaur (36), Scattered (28), and Detached (15) objects.

This sample shows a bimodal distribution of optical spectral slopes (s); a Gaussian mixture model fit to the histogram indicated that the two colour classes intersect at $s \simeq 20.6\% \, (10^3)^{-1}$, corresponding to $(g-r) \approx 0.78$ and $(V-R) \approx 0.56$. This value agrees with thresholds adopted in earlier works (see Marsset et al. 2019, Table 2) and provides a boundary between less-red and very-red objects: TNOs with $s + \delta s < 20.6\% \, (10^3)^{-1}$ were labelled ‘gray/LROs,’ (i.e., less-red objects) those with $s - \delta s > 20.6\% \, (10^3)^{-1}$ were labelled ‘very-red/VROs,’ (i.e., very-red objects) and objects whose uncertainties straddle the boundary were left unclassified and out of the dataset. Using this classification, Marsset et al. (2019) found that the very-red population has a cut-off inclination of $\sim 21°$, whereas gray objects extend to higher inclinations. Their results were subsequently expanded by Ali-Dib et al. (2021), who reported a similar cut-off in eccentricity around $e \approx 0.42$ for the VROs. Ali-Dib et al. (2021) concluded that there is a paucity of VROs in the scattered disk and used a Solar System formation model to explain these trends as a consequence of their formation location in the disk.

Our dataset is further summarised in Fig. 2. We define VROs as TNOs with spectral slopes greater than $20.6\%/(10^3 \text{Å})$. The colour–eccentricity correlation is revealed in this plot as a paucity of VROs for eccentricity above 0.42. Similarly, the colour–inclination correlation manifests itself as a lack of VROs for inclinations above 21°.

2.2 The Dark Energy Survey

The Dark Energy Survey (Bernardinelli et al. 2023, 2025) sample consists of 814 TNOs with absolute magnitude $5.5 < H_r < 8.2$ and accurate colours. The obtained optical colour classes were found to be an identical match to the different compositional classes derived from *JWST* IR spectra (when available). For self-consistency between the two datasets, we removed cold objects with inclinations below 5°, and very far objects beyond the maximal semimajor axis of the Col-OSSOS dataset. We thus end up with 674 TNOs spanning most dynamical classes: 274 Hot Classicals, 209 Resonant, 145 Detached,

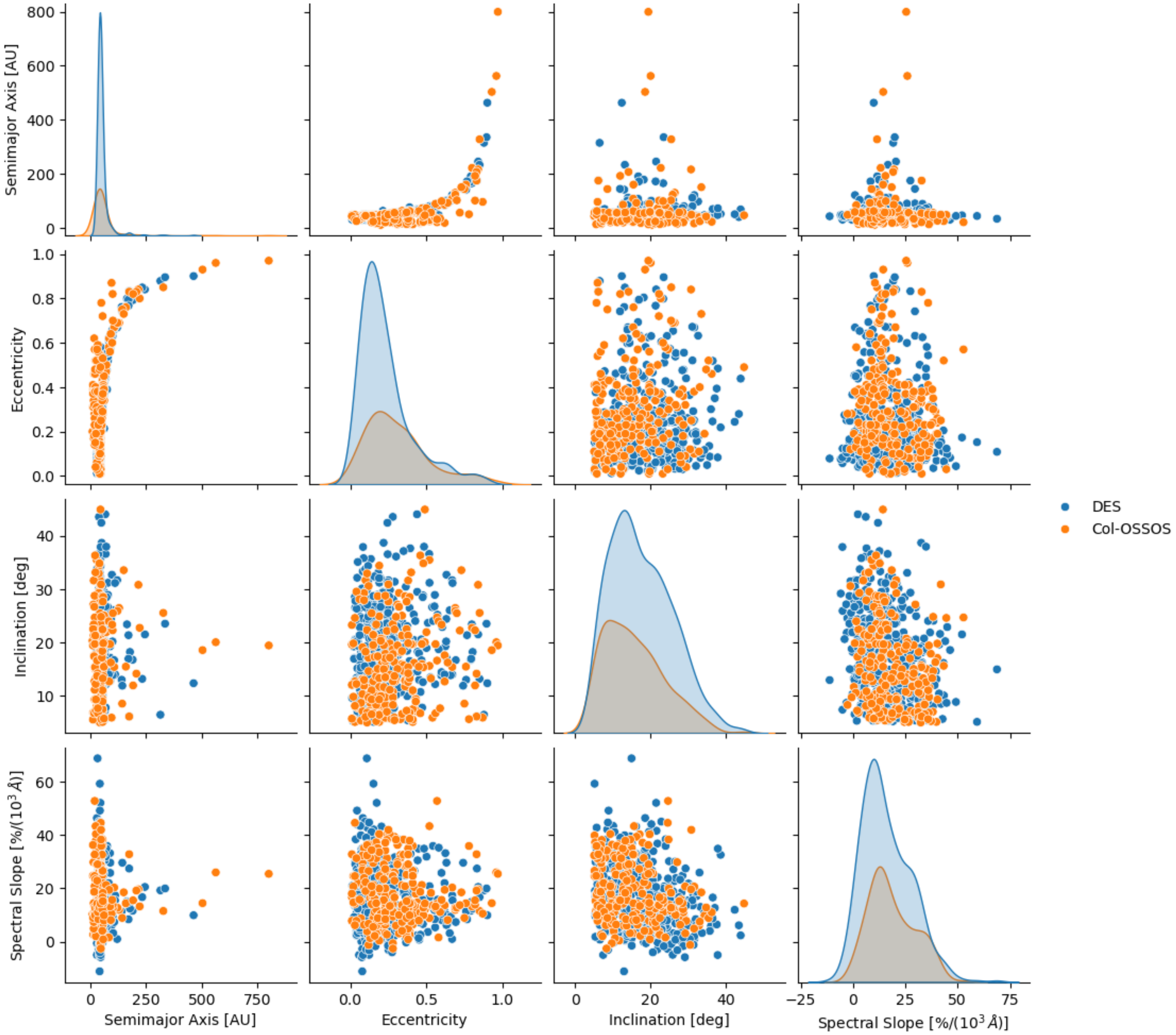

Figure 1. Pairplot of all 229 and 674 TNOs in the Col-OSSOS and DES datasets, respectively. The colour–eccentricity and colour–inclination correlations hold in both datasets.

45 Scattered, and one Centaur.[1] Our final datasets are shown in Fig. 1, where the correlations between colour and inclination, and colour and eccentricity are noticeable in both samples, confirming that they are not merely an observational bias in the Col-OSSOS sample.[2]

3 METHODOLOGY: CAUSAL DISCOVERY

The foundation of causal discovery lies in uncovering the footprints of causality embedded in data. One of the most important sources of such information is conditional independency (CI) relations. In this section, we will first show how different causal structures leaves different CI footprints, then explain how to utilise these CIs to constrain causal structures even with the presence of latent variables, and finally we introduce the Fast Causal Inference (FCI) algorithm used in this work. For further reading on causal discovery and causality, see *Causation, Prediction, and Search* (Spirtes et al. 2001), *Causality* (Pearl 2009), or a review for astrophysicists in Jin et al. (2025b, §2).

[1] Note that centaurs are much less prominent than in the Col-OSSOS sample. Thus, the more unbalanced DES sample is less suited for demographic studies between the subpopulations.

[2] Note that Bernardinelli et al. (2025), used this survey to find a paucity of 'near-infrared faint' SDOs (scattered-disk objects) in the data.

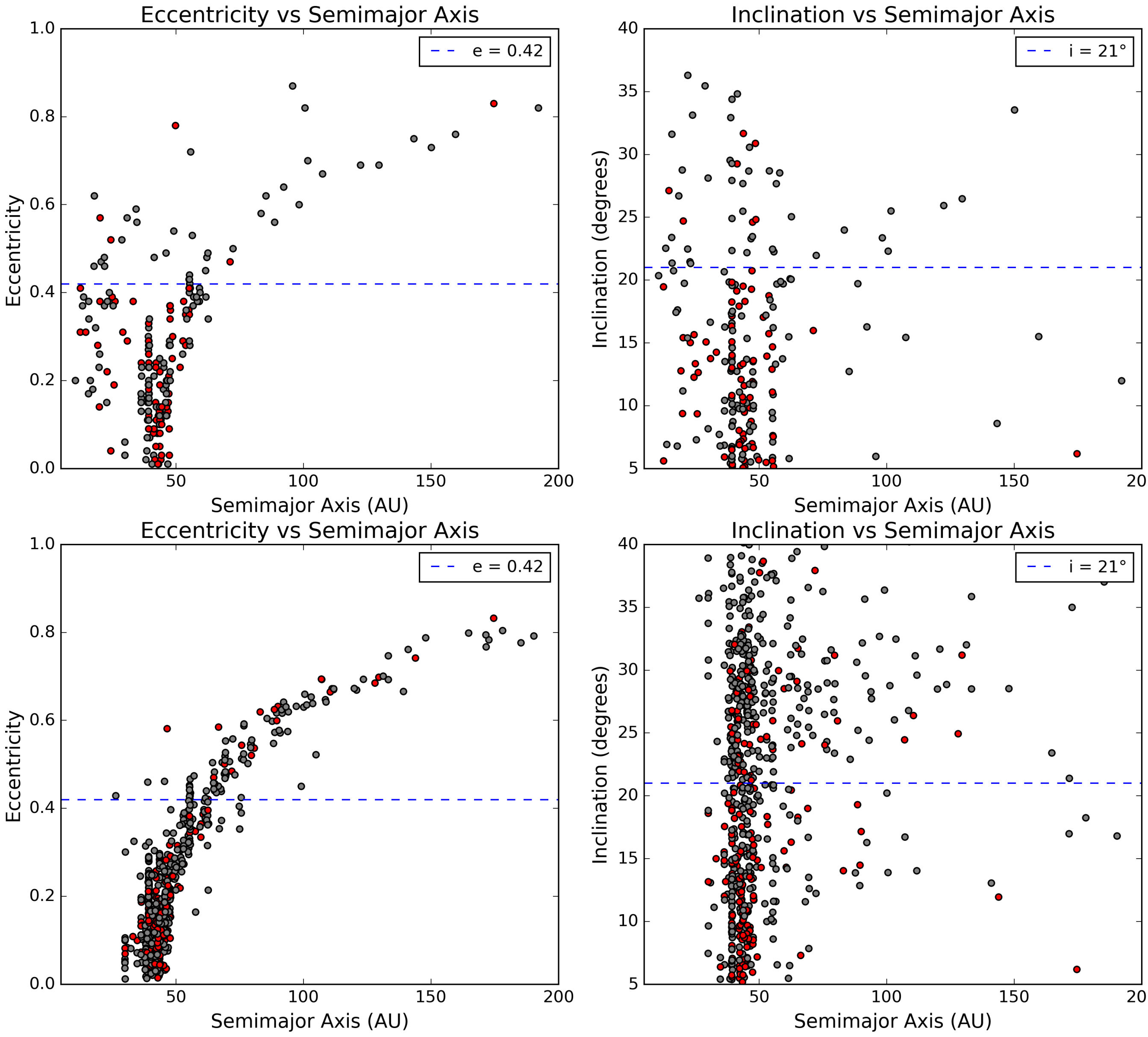

Figure 2. The Marsset et al. (2019) and Ali-Dib et al. (2021) (top), and DES (bottom) samples shown as a–e (left) and a–inc (right) plots. Colours were defined such that red (●) is for Very Red Objects (spectral slopes higher than 20.6%/(10^3Å)) and gray (●) is for Less Red Objects. The plots clearly show the paucity of VROs for eccentricities higher than 0.42 and inclinations higher than 21°, respectively (- - -).

3.1 Conditional independency footprints

The causal structure among a set of variables is ideally represented by a Directed Acyclic Graph (DAG), consisting of nodes and directed edges (i.e., arrows), where directed edges between nodes denote the direction of causality. There are three basic causal structures: *chains*, *forks*, and *colliders* (Figure 3), each of them carrying two (conditional) independency signatures. These sets of CIs naturally hold given each DAG, as long as a rather general assumption called the Markov assumption[3] holds. It is worth noting that although a chain and a fork share the same CIs, a collider carries a different set of CIs, making it possible to constrain causal structures. The three basic causal structure serve as the building block towards more composite DAGs with more than three variables. The CIs in these larger DAGs are determined by all the paths among variables, where each path is made of chains, forks, or colliders. The CIs of chains, forks, and colliders propagate along the paths following separation rules, which we will not go into detail here.

3.2 Causal discovery with latent variables

By analysing conditional independencies among different components of an observed system, we can infer causal relationships between pairs of variables. This allows us to construct a graph that encodes the results of essential conditional independence tests, re-

[3] The Markov assumption states that given its parents in a DAG, a node is independent of all its non-descendants.

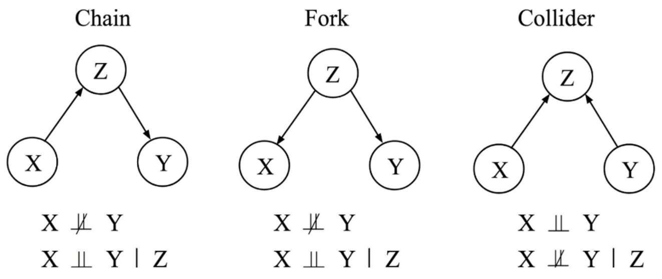

Figure 3. Three basic causal blocks, and their resulting set of (conditional) independencies. $\not\perp\!\!\!\perp$ denotes dependent, $\perp\!\!\!\perp$ means independent, and | is the notation for condition. For example, '$X \perp\!\!\!\perp Y \mid Z$' means '$X$ is independent to Y when conditioned on Z.'

vealing which variables *cause* others under appropriate conditions. Ideally, the output is a DAG for a unique solution (for example, in the case of a collider) or a Completed Partially Directed Acyclic Graph (CPDAG) for a Markov equivalence class, where the direction of some of the edges cannot be determined (for example, in the case of a chain or a fork).

However, since it is impossible to measure all variables in the Universe, latent variables are always present. These unmeasured variables can significantly impact the correctness of the causal structure discovered. For example, suppose that X and Y are independent in the general population, but a sample is selected based on a variable Z that influences both X and Y. In that case, X and Y may exhibit statistical dependence in the sample, even though no such relationship exists in the population. This can lead to spurious causal conclusions, falsely suggesting a direct causal relationship between X and Y.

To address this challenge, we employ a principled approach capable of uncovering causal relationships even in the presence of latent variables. A widely used method for this purpose is Fast Causal Inference (FCI; Spirtes et al. 1995; Zhang 2008), a constraint-based algorithm that has been proven to provide sound causal conclusions despite unmeasured variables. FCI has been applied across various scientific domains, including biology, economics, and climate science. For our analysis of TNO orbital elements and colours, we use the FCI implementation in the `Python` package `causal-learn` (Zheng et al. 2024) to infer the underlying causal structure.

3.3 The Fast Causal Inference algorithm

Unlike many causal discovery methods that assume that all relevant variables are measured (such as those producing DAGs or CPDAGs), FCI accounts for the possibility of unobserved variables. As a result, its output is a Partial Ancestral Graph (PAG), which provides more nuanced causal information.

3.3.1 Partial Ancestral Graph notations

In a PAG, a bi-directional arrow $X \longleftrightarrow Y$ corresponds to a confounding relation (i.e., a third variable causes both X and Y) and empty circles ($\circ$) represent uncertainty regarding the ending symbol of an edge. Specifically, the edges in a PAG have the following interpretations:

- $X \longrightarrow Y$: X is a *cause* of Y.
- $X \longleftrightarrow Y$: There is a latent common cause of X and Y.
- $X \circ\!\!\longrightarrow Y$: Y is not an *ancestor* of X, i.e., either $X \rightarrow Y$ or $X \longleftrightarrow Y$, but not $X \leftarrow Y$.
- $X \circ\!\!-\!\!\circ Y$: No set d-separates X and Y, i.e., $X \rightarrow Y$, $X \leftarrow Y$, and $X \longleftrightarrow Y$ are all possible.

By introducing these notations, we can account for latent variables in the discovery process and uncover causal relations among measured variables while acknowledging uncertainties introduced by unmeasured factors. More importantly, when the algorithm cannot determine a definitive causal direction due to latent variables, it explicitly represents this uncertainty rather than arbitrarily assigning a direction. This principled approach distinguishes causal analysis from correlation-based techniques, ensuring that conclusions are drawn with a clear acknowledgment of underlying assumptions and limitations.

3.3.2 FCI: A two-stage algorithm based on CI tests

FCI is a two-stage algorithm to discovers causal relationships based on a series of CI tests. The idea is to start with a fully-connected graph[4], where all causal structures are allowed, and gradually constrain the graph when the CIs encoded by the graph are inconsistent with the CIs found in the data. Specifically, the algorithm proceeds in the follow two stages:

(i) *Skeleton Discovery*: Starting with a fully connected graph representing all possible causal structures, FCI removes edges when two nodes are independent, and when two nodes become conditionally independent given some subset of other variables. By the end of this stage, FCI arrives at an the undirected graph called a 'skeleton.'

(ii) *Orientation*: The second stage is to orient the remaining edges. FCI orients the edges based on the separation information, collider detection (e.g., identifying V-structures), and propagation of orientation constraints.

We will walk readers through the above two FCI stages in our specific TNO case in §3.3.4.

3.3.3 Conditional independence tests

CI tests are performed along with the two stages of FCI (mostly during stage i, and the results are re-used during stage ii) whenever necessary. Here, we adopt two commonly used CI tests in causal discovery, including the Fisher Z-test (Fisher 1921) and the Kernel-based Conditional Independence (KCI) test (Zhang et al. 2011).

- The Fisher Z-test assesses conditional independence by measuring partial linear correlations between variables, providing a fast and effective method for detecting dependence relations.
- In contrast, the KCI test is a non-parametric approach based on reproducing kernel Hilbert space (RKHS) embeddings, allowing it to capture complex non-linear dependencies without assuming specific functional forms.

Like many other statistical tests, the output of the Fisher Z-test or the KCI test is a p-value which gives the possibility against a null (conditionally dependent) hypothesis. For example, if the CI test between A and B conditioned on both C and D has a p-value of 0.05, then $A \not\perp\!\!\!\perp B \mid \{C, D\}$ at the 95% confidence level.

While KCI offers greater flexibility and consistency in general settings, it is computationally more intensive and sensitive to kernel

[4] In a fully-connected graph, every node is connected to every other node with the most general type of edge $\circ\!\!-\!\!\circ$, allowing any type of potential causal connection between any pair of nodes.

choices. The selection of CI test depends on the trade-off between computational efficiency and the need for non-parametric estimation. Here, we perform both the Fisher Z-test and the KCI test with polynomial kernels to both Col-OSSOS and DES datasets. The CIs and p-values are reported in Table 1. For any dataset and any CI test (i.e., any of the four columns), $a \perp\!\!\!\perp$ inc, $a \perp\!\!\!\perp$ colour, and $e \perp\!\!\!\perp$ colour show significant higher p-values compared to others within the same column, indicating they are independent, while other CIs listed are dependent. Quantitatively, the CIs are determined at least above the 98.2% confidence level for Col-OSSOS with the Fisher Z-test, 91.5% for Col-OSSOS with the KCI test, 99.9% for DES with the Fisher Z-test, and 99.7% for DES with the KCI test.[5]

3.3.4 Step-by-step derivation of the PAG

Here, we manually derive of the causal structure (i.e., the PAG) following the two stages of FCI outlined in §3.3.2 using the list of CIs discussed in the previous section (§3.3.3) and shown in Table 1. The whole process is automated in the `Python` package `causal-learn` (Zheng et al. 2024), but here we explicitly write it out for readers new to causal discovery.

During FCI stage i, we start from the hypothesised, undirected, and fully-connected graph in Fig. 4 (top panel) with our four nodes (a, e, inc, and colour), and remove edges between independent nodes. For example, given $a \perp\!\!\!\perp$ inc, it is then not possible to have a cause inc ($a \longrightarrow$ inc), inc cause a (inc $\longrightarrow a$), nor a third latent variable (L) cause both a and inc ($a \longleftarrow L \longrightarrow$ inc, shortened as $a \leftrightarrow$ inc). Therefore, the edge between a and inc (a ∘—∘ inc) can be removed. Similarly, a ∘—∘ colour and e ∘—∘ colour are also removed since $a \perp\!\!\!\perp$ colour and $e \perp\!\!\!\perp$ colour. The remaining edges are valid since the nodes are dependent with or without conditioning on other non-latent nodes. For example, $e \not\!\perp\!\!\!\perp$ inc, $e \not\!\perp\!\!\!\perp$ inc $\mid a$, and $e \not\!\perp\!\!\!\perp$ inc $\mid$ colour together requires an edge between e and inc.[6]

Now, we are left with with a skeleton graph as Fig. 4 (bottom panel), and we can move to FCI stage ii—orient the remaining edges according to CIs. Given the current skeleton, $a \perp\!\!\!\perp$ colour, $a \perp\!\!\!\perp$ inc, and $e \perp\!\!\!\perp$ colour together form a classical setup where there must be a latent confounder between e and inc. Consider a, e, and inc, both a chain structure (i.e., a∘⟶ e∘⟶ inc or a ⟵∘e ⟵∘inc) and a fork structure (i.e., a ⟵∘e∘⟶ inc) are forbidden as they will not satisfy the fact that $a \perp\!\!\!\perp$ inc. The only structure compatible with $a \perp\!\!\!\perp$ inc is a collider (i.e., a∘⟶ e ⟵∘inc). Similarly, we can find e∘⟶ inc ⟵∘colour according to $e \perp\!\!\!\perp$ colour. The need for both e ⟵∘inc and e∘⟶ inc calls for a latent confounder L causing both e and inc (i.e., $e \longleftarrow L \longrightarrow$ inc, or $e \leftrightarrow$ inc in a more compact notation). We therefore arrive at the final PAG in Fig. 4 (bottom panel) and later shown in Fig. 5.

3.3.5 Validation of FCI with generated data

The FCI algorithm is a time-tested algorithm that has been proven successful both in idealised data (Colombo et al. 2012) and real-world data (Glymour et al. 2019). Here, we perform a simple test with generated ideal data from latent linear Structural Causal Models (SCMs). Such models can be defined as a DAG $\mathcal{G} := (\mathbf{V}_{\mathcal{G}}, \mathbf{E}_{\mathcal{G}})$, where each variable $V_i \in \mathbf{V}_{\mathcal{G}}$ is generated following a latent linear SCM:

$$\mathsf{V}_i = \sum\nolimits_{\mathsf{V}_j \in \mathrm{Pa}_{\mathcal{G}}(\mathsf{V}_i)} a_{ij}\mathsf{V}_j + \varepsilon_{\mathsf{V}_i}, \tag{1}$$

where $\mathbf{V}_{\mathcal{G}} := \mathbf{L}_{\mathcal{G}} \cup \mathbf{X}_{\mathcal{G}}$ contains a set of n observed variables ($\mathbf{X}_{\mathcal{G}} := \{\mathsf{X}_i\}_{i=1}^{n}$) and m latent variables ($\mathbf{L}_{\mathcal{G}} := \{\mathsf{L}_i\}_{i=1}^{m}$). $\mathrm{Pa}_{\mathcal{G}}(V_i)$ is the parent set (i.e., nodes that directly cause V_i) of V_i, a_{ij} denotes the causal coefficient from V_j to V_i, and $\varepsilon_{\mathsf{V}_i}$ represents the noise term. Following this latent linear SCM setup, we generate multiple mock datasets with a random DAG containing both observed variables and latent variables, $\varepsilon_{\mathsf{V}_i}$, randomly sampled from Gaussian distributions with a random mean and standard deviation $N(\mu_i, \sigma_i)$, and a random value of a_{ij}. We apply the FCI algorithm to only observed variables, and we find the FCI algorithm is able to uncover the correct PAG corresponding to the ground-truth DAG in all of the generated datasets.

4 RESULTS

Our primary finding is the causal structure found among a, e, inc, and colour, shown as a PAG in Fig. 5. The notations of the PAG can be found in §3.3.1, and the PAG is derived through the FCI algorithm outlined in §3.3.2 and detailed in §3.3.4. The PAG is based on the result of a set of CIs shown in §3.3.3 and Table 1. These CIs have confidence levels of at least 98.2% and 91.5% for Col-OSSOS data according to two different CI tests, and confidence levels of at least 99.9% and 99.7% for DES with the two CI tests. Moreover, we consistently reproduce the same PAG as in Fig. 5 by jackknifing our data by sequentially leaving out each subpopulation of TNOs. Thus, removing any subsample of 48 Classicals, 102 Resonant, 36 Centaurs, 28 Scattered, or 15 Detached TNOs among the Col-OSSOS data results in no change to our discovered PAG. Therefore, we demonstrate that no single subpopulation is dominating the PAG and that our results are robust to outliers.

Alternatively, if we are to generate PAGs for the individual populations separately (i.e., analysing only one subpopulation at a time), we find a large diversity in the results. Many of these PAGs, however, are based on very few data points. Taking this result at face value hints that our overall PAG represents that main-line dynamics dominate over the entire sample. We emphasize that this PAG was obtained with a purely data-driven approach, without astrophysical foresight.

We get a similar result when following the same procedure for the DES sample, except when removing the detached objects population where we end up with a slightly different PAG. However, running the model on the detached objects alone also leads to the same PAG, implying that while they contribute significantly to the result, they are not solely responsible for it as a subpopulation. The main difference in CIs that leads to this different PAG is the independence between e and inc both with and without detached objects. As the detached objects contribute to the high-eccentricity–high-inclination subpopulation while scattered objects have high eccentricity and low inclination, removing detached objects dilutes the global correlation between e and inc. The detached subpopulation itself moreover does not have a strong correlation between the two since they do not undergo significant von Zeipel-Lidov-Kozai oscillations (von Zeipel 1910; Lidov 1962; Kozai 1962).

[5] There are more data points in DES (674) than in Col-OSSOS (229), therefore the CI tests on DES data show lower p-values (i.e., higher confidence levels) than the ones on Col-OSSOS data, as expected.

[6] $e \not\!\perp\!\!\!\perp$ inc alone does not guarantee an edge between e and inc, as $e \longleftarrow a \longrightarrow$ inc or $e \longleftarrow$ colour $\longrightarrow$ inc can both lead to $e \not\!\perp\!\!\!\perp$ inc. However $e \not\!\perp\!\!\!\perp$ inc $\mid a$ and $e \not\!\perp\!\!\!\perp$ inc $\mid$ colour rules out these two cases and secures an edge between e and inc.

Table 1. List of conditional independencies, type of conditional independence tests performed, and the p-value of the statistical test. $\not\perp\!\!\!\perp$ denotes dependent, $\perp\!\!\!\perp$ means independent, and $|$ is the notation for condition. The Fisher-Z and KCI tests are performed on both Col-OSSOS and DES data. The p-value for the null hypothesis for each CI test is shown. A p-value closer to zero suggests dependence, and a p-value closer to one favours independence. Both tests on both datasets unanimously show that $a \perp\!\!\!\perp$ inc, $a \perp\!\!\!\perp$ colour, and $e \perp\!\!\!\perp$ colour, while other CIs listed are dependent. The PAG derived in Fig. 4 and later shown in Fig. 5 directly comes from this list of conditional independencies following the FCI algorithm detailed in §3.3.2 and §3.3.4.

Conditional independencies	Conditional independence test p-value			
	Col-OSSOS		DES	
	Fisher-Z	KCI	Fisher-Z	KCI
$a \not\perp\!\!\!\perp e$	0.000	0.000	0.000	0.000
$a \perp\!\!\!\perp$ inc	0.144	0.421	0.006	0.568
$a \perp\!\!\!\perp$ colour	0.918	0.488	0.677	0.250
$e \not\perp\!\!\!\perp$ inc	0.005	0.085	0.000	0.001
$e \perp\!\!\!\perp$ colour	0.211	0.135	0.181	0.010
inc $\not\perp\!\!\!\perp$ colour	0.001	0.009	0.000	0.000
$a \not\perp\!\!\!\perp e \mid$ inc	0.000	0.068	0.000	0.001
$e \not\perp\!\!\!\perp$ inc $\mid a$	0.018	0.070	0.000	0.002
$e \not\perp\!\!\!\perp$ inc $\mid$ colour	0.010	0.041	0.000	0.003
inc $\not\perp\!\!\!\perp$ colour $\mid e$	0.003	0.004	0.000	0.000

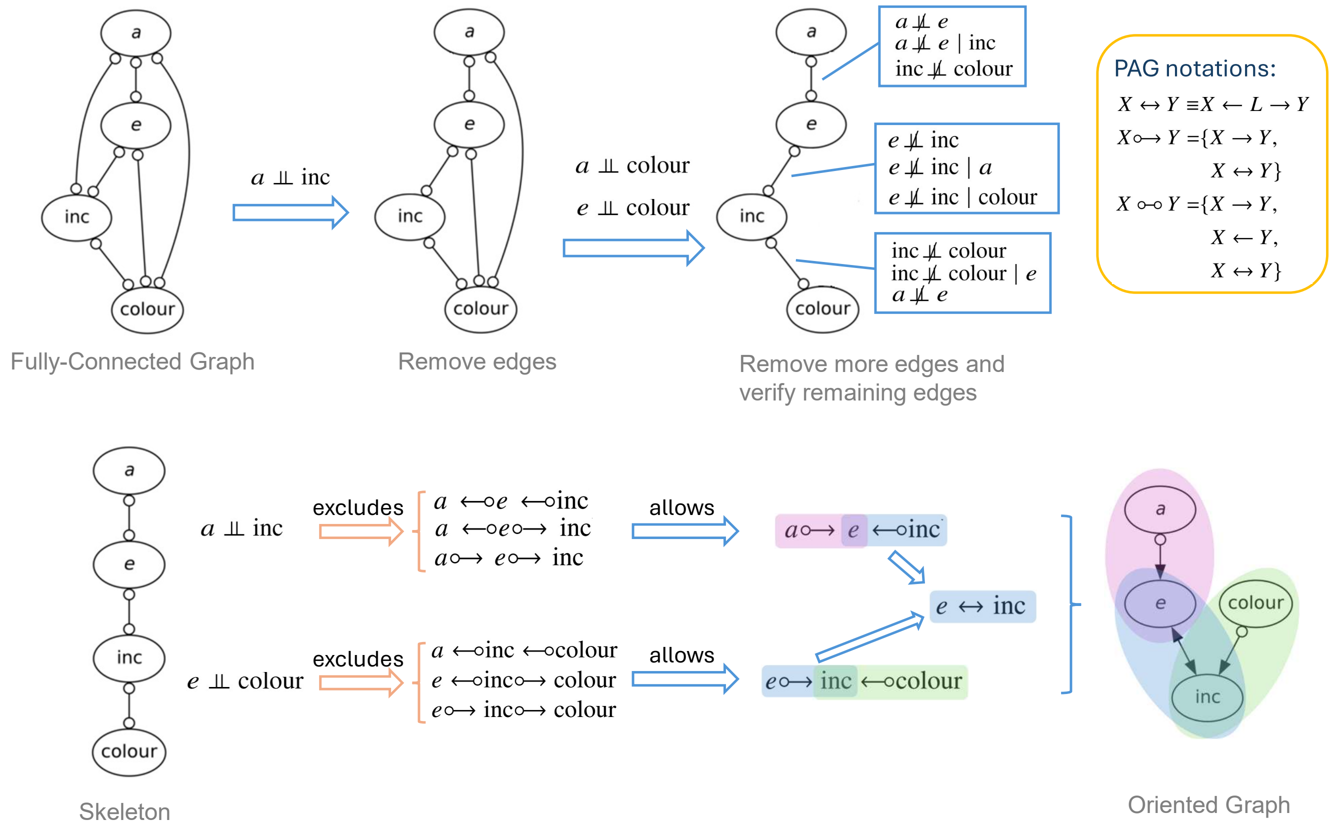
a
e
inc
colour
Fully-Connected Graph
$a \perp\!\!\!\perp$ inc
Remove edges
$a \perp\!\!\!\perp$ colour
$e \perp\!\!\!\perp$ colour
$a \not\perp\!\!\!\perp e$
$a \not\perp\!\!\!\perp e \mid$ inc
inc $\not\perp\!\!\!\perp$ colour
$e \not\perp\!\!\!\perp$ inc
$e \not\perp\!\!\!\perp$ inc $\mid a$
$e \not\perp\!\!\!\perp$ inc $\mid$ colour
inc $\not\perp\!\!\!\perp$ colour
inc $\not\perp\!\!\!\perp$ colour $\mid e$
$a \not\perp\!\!\!\perp e$
Remove more edges and verify remaining edges
PAG notations:
$X \leftrightarrow Y \equiv X \leftarrow L \rightarrow Y$
$X \circ\!\!\rightarrow Y = \{X \rightarrow Y, X \leftrightarrow Y\}$
$X \circ\!\!-\!\!\circ Y = \{X \rightarrow Y, X \leftarrow Y, X \leftrightarrow Y\}$
Skeleton
$a \perp\!\!\!\perp$ inc
excludes
$a \leftarrow\!\!\circ e \leftarrow\!\!\circ$ inc
$a \leftarrow\!\!\circ e \circ\!\!\rightarrow$ inc
$a \circ\!\!\rightarrow e \circ\!\!\rightarrow$ inc
allows
$a \circ\!\!\rightarrow e \leftarrow\!\!\circ$ inc
$e \leftrightarrow$ inc
$e \perp\!\!\!\perp$ colour
excludes
$a \leftarrow\!\!\circ$ inc $\leftarrow\!\!\circ$ colour
$e \leftarrow\!\!\circ$ inc $\circ\!\!\rightarrow$ colour
$e \circ\!\!\rightarrow$ inc $\circ\!\!\rightarrow$ colour
allows
$e \circ\!\!\rightarrow$ inc $\leftarrow\!\!\circ$ colour
Oriented Graph

Figure 4. The visualisation of the FCI algorithm. In FCI stage i (upper panel), the algorithm starts with a fully-connected graph where all nodes are interconnected with ∘—∘ to allow all possible cases. Then, the algorithm goes through every edge and removes an edge when two nodes are independent, and when two nodes become conditionally independent given some subset of other variables. In FCI stage ii (lower panel), the remaining edges are oriented by constraining the ending symbols of each edge according to conditional independencies.

4.1 Astrophysical interpretation

In this subsection, we investigate whether the causal links found by our model are consistent with the physical mechanisms at play in the Kuiper belt. We emphasise that that our method is not capable of re-discovering these mechanisms or deriving physical laws from data, but should simply be compatible with them.

4.1.1 $a \circ\!\!\rightarrow e$

The first link we investigate is the one-way causal direction of the *current* semimajor axis causing the *current* eccentricity. While the correlation between a and e in TNOs is well established, the direction of the causality we find here is not surprising either, as its root physical causes are:

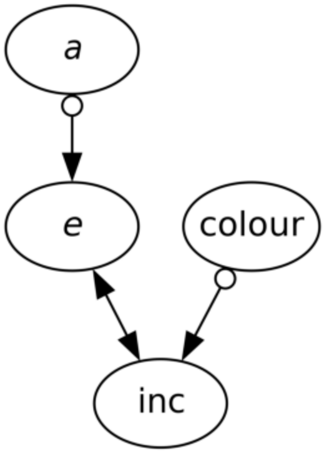

Figure 5. Partial Ancestral Graph (PAG) calculated with the Fast Causal Inference (FCI) algorithm (Spirtes 2001; Spirtes et al. 2013; Zheng et al. 2024) with linear Fisher-Z conditional independence tests (Fisher 1921) as well as non-linear Kernel-based conditional independence (KCI) tests (Zhang et al. 2012). This PAG has three causal edges, which can be described as follows: (i) eccentricity is not an ancestor of the semimajor axis, (ii) there is a latent common cause of eccentricity and inclination, and (iii) inclination is not an ancestor of colour.

- Scattering by Neptune, where objects have to close-encounter Neptune first in order to get scattered into high-eccentricity orbits. Moreover, objects usually cannot be both close to Neptune (today) and have a high eccentricity. It is the current semimajor axis of the objects that dictates what eccentricity they can have, and not the other way around.
- Mean motion resonances (MMRs), where the period (and thus current semimajor axis) of the objects dictates whether they are inside an eccentricity-raising resonance.

The connection $a \circ\!\!\rightarrow e$ *rules out* the possibility of $a \leftarrow e$. Clearly, $a \rightarrow e$ is possible, but also $a \leftrightarrow e$. The latter might imply that an unobserved confounder causes both a and e.

4.1.2 $e \leftrightarrow$ inc

The second link in the PAG is the two-way dependency between the eccentricity and the inclination, which is consistent with the von Zeipel-Lidov-Kozai anti-correlated oscillations between these two quantities (both inside and outside of MMRs), that plays a central role in the dynamics of TNOs. Here, $e \leftrightarrow$ inc implies that there is an unobserved confounder. Indeed, the von Zeipel-Lidov-Kozai mechanism involves perturbations from a third body, here being Neptune. Moreover, if Neptune had not already been predicted in 1821 and eventually identified in 1846, our result here would strongly suggest the presence of an unknown perturbing body. Together, the first two links successfully re-establish the main dynamical processes shaping the Kuiper belt (scattering, MMRs, and von Zeipel-Lidov-Kozai oscillations) without any physical inputs.

4.1.3 colour $\circ\!\!\rightarrow$ inc

Finally, the third piece of the puzzle is the connection colour $\circ\!\!\rightarrow$ inc, *ruling out* the possibility of colour ← inc. The 'colour' (i.e., a proxy for the formation location in our null hypothesis) is hence consistent with causing the inclination (colour → inc). This is again dynamically expected, as the formation location relative to inclination-raising secular resonances, such as f_7 and f_8, will strongly affect the inclination distribution of TNOs (Murray & Dermott 1999). However, this link leaves open the possibility of an unobserved confounder causing both colour and the inclination (colour ↔ inc). This confounder can be the formation location itself, if we were to assume the colour and initial location to be two distinct variables instead of the colour being a proxy for location.

Our result, that colour ← inc is not allowed, rules out the model of Luu & Jewitt (1996) and Stern (2002), where collisional evolution shapes the colours of TNOs. Moreover, our result that colour ← a is not allowed either, rules out models based solely on Kaňuchová et al. (2012), where the current a would control the amount of irradiation a TNO is subjected to, and thus its colour. However, our findings cannot confirm or exclude the pre-instability irradiation scenario proposed by Brown et al. (2011) and Wong & Brown (2016, 2017), as our analysis cannot disentangle the effect of formation location from immediate surface modification at the location of formation. These findings are consistent with the recent results of Belyakov et al. (2024) and Licandro et al. (2025).

4.1.4 Further interesting features found in the PAG

- *The lack of correlation between the colour and semimajor axis*: This is dynamically expected as all TNOs in our sample underwent dynamical interactions with Neptune, that tend to be chaotic in nature. For example, many of the relevant processes (scattering, resonances, etc.) depend on the phase angle at which the TNO encounters Neptune. Some examples of the chaotic outcomes of the TNO dynamics are shown in Ali-Dib et al. (2021, Figs. 11 and 12). See also Nesvorný et al. (2016, Fig. 3).
- *The indirect causation between the colour (initial location) and the eccentricity via the inclination*: Taken at face value, this would indicate that while the initial location directly causes the inclination, it is the final semimajor axis that causes the eccentricity. The effect of the initial semimajor axis on the eccentricity is indirect, and happens through von Zeipel-Lidov-Kozai oscillations starting from high inclinations.

5 DISCUSSION & CONCLUSIONS

Our work endeavours to resolve the tension between theories of primordial origins vs. subsequent evolution to account for the observed dispersion and correlations in TNO colours, a subject of a long debate. Our causal graph analysis, derived from a model-agnostic causal discovery framework, strongly favours the primordial origin hypothesis that *TNO colour is causally antecedent to inclination, not a consequence of it.* Accordingly, we have high confidence (>91.5%–99.9%) in our result because it is unanimously found from both the Col-OSSOS and DES samples, with each dataset analysed by Fisher-Z and KCI tests. While impacts undoubtedly modify surfaces, our results suggest they are not the dominant driver of colour diversity. Moreover, our model seems to exclude any effects from the current semimajor axis on the colour of TNOs, disfavouring models where continuous irradiation plays a large role in shaping the colours. The consequences of having a colour gradient in the outer protoplanetary disk, as implied by this work (and many others as discussed), open up new possibilities to resolve the the Trojan colour conundrum (Jewitt 2018). This will be explored further in future work.

While many earlier works tried to explain the inclination–colour and eccentricity–colour correlations, both separately and simultaneously, our causal approach isolates inclination as the key dynamical variable causally linked to colour. This hints at a larger role for inclination-raising secular resonances in the very early Solar System. Indeed, Ali-Dib et al. (2021) proposed that the origins of the paucity of VROs in the scattered disk are strongly linked to the f_7 and f_8 inclination modes. In this scenario, the colour–eccentricity correlation is largely (although not necessarily entirely) a consequence of the more fundamental inclination–colour correlation, where the two can be linked via the von Zeipel-Lidov-Kozai mechanism. This is consistent with the numerical model of Ali-Dib et al. (2021), who proposed von Zeipel-Lidov-Kozai oscillations as a transport vehicle for VROs between high-inclination and high-eccentricity regimes.

Finally, our work is a proof of principle for the use of causality models in planetary sciences. Moreover, our results will be a valuable addition to the study of Kuiper Belt Object colours with new data coming from the Vera C. Rubin Observatory.

ACKNOWLEDGEMENTS

We thank the anonymous referee for their insightful comments that helped improve the clarity and quality of this manuscript. This material is based on work supported by Tamkeen under the NYU Abu Dhabi Research Institute grant CASS. YZ and KZ are supported by NSF Award No. 2229881, AI Institute for Societal Decision Making, NIH R01HL159805, and grants from Quris AI, Florin Court Capital, and MBZUAI-WIS Joint Program. This research was carried out on the high-performance computing resources at New York University Abu Dhabi. This research has made use of NASA's Astrophysics Data System Bibliographic Services.

DATA AVAILABILITY

The data and code used for this work are available for download from the following GitHub repository: `https://github.com/ZehaoJin/causalTNOs`.

REFERENCES

Ali-Dib M., Marsset M., Wong W.-C., Dbouk R., 2021, AJ, 162, 19
Ayala-Loera C., Alvarez-Candal A., Ortiz J. L., Duffard R., Fernández-Valenzuela E., Santos-Sanz P., Morales N., 2018, MNRAS, 481, 1848
Belyakov M., Brown M. E., Al-Kibbi A., 2024, PSJ, 5, 193
Bernardinelli P. H., et al., 2023, ApJS, 269, 18
Bernardinelli P. H., et al., 2025, AJ, 169, 305
Brown M. E., 2012, Annual Review of Earth and Planetary Sciences, 40, 467
Brown M. E., Schaller E. L., Fraser W. C., 2011, ApJ, 739, L60
Buchanan L. E., et al., 2022, PSJ, 3, 9
Chen Y.-T., Eduardo M. R., Muñoz-Gutiérrez M. A., Wang S.-Y., Lehner M. J., Chang C.-K., 2022, ApJ, 937, L22
Colombo D., Maathuis M. H., Kalisch M., Richardson T. S., 2012, The Annals of Statistics, pp 294–321
Davis B. L., Ali-Dib M., Zheng Y., Jin Z., Zhang K., Macciò A. V., 2025, arXiv e-prints, p. arXiv:2507.03760
Fisher R. A., 1921, Metron, 1, 3
Friedman N., 2004, Science, 303, 799
Glymour C., Zhang K., Spirtes P., 2019, Frontiers in genetics, 10, 524
Hainaut O. R., Delsanti A. C., 2002, A&A, 389, 641
Jewitt D., 2018, AJ, 155, 56
Jewitt D. C., Luu J. X., 2001, AJ, 122, 2099
Jin Z., Pasquato M., Davis B. L., Macciò A. V., Hezaveh Y., 2024, arXiv e-prints, p. arXiv:2410.14775
Jin Z., Pasquato M., Davis B., Maccio A., Hezaveh Y., 2025a, in American Astronomical Society Meeting Abstracts. p. 120.03D
Jin Z., et al., 2025b, ApJ, 979, 212
Kaňuchová Z., Brunetto R., Melita M., Strazzulla G., 2012, Icarus, 221, 12
Kozai Y., 1962, AJ, 67, 591
Licandro J., et al., 2025, Nature Astronomy, 9, 245
Lidov M. L., 1962, Planet. Space Sci., 9, 719
Liu P.-Y., Ip W.-H., 2019, ApJ, 880, 71
Luu J. X., Jewitt D. C., 1996, AJ, 112, 2310
Marsset M., et al., 2019, AJ, 157, 94
Morbidelli A., Nesvorný D., 2020, in Prialnik D., Barucci M. A., Young L., eds, , The Trans-Neptunian Solar System. pp 25–59, doi:10.1016/B978-0-12-816490-7.00002-3
Murray C. D., Dermott S. F., 1999, Solar System Dynamics, doi:10.1017/CBO9781139174817.
Nesvorný D., Vokrouhlický D., Roig F., 2016, ApJ, 827, L35
Nesvorný D., et al., 2020, AJ, 160, 46
Pasquato M., 2024, in EAS2024, European Astronomical Society Annual Meeting. p. 362
Pasquato M., Jin Z., Lemos P., Davis B. L., Macciò A. V., 2023, arXiv e-prints, p. arXiv:2311.15160
Pearl J., 2009, Causality. Cambridge university press
Pinilla-Alonso N., et al., 2025, Nature Astronomy, 9, 230
Runge J., et al., 2019, Nature Communications, 10, 2553
Sachs K., Perez O., Pe'er D., Lauffenburger D. A., Nolan G. P., 2005, Science, 308, 523
Schwamb M. E., et al., 2019, ApJS, 243, 12
Spirtes P., 2001, in Richardson T. S., Jaakkola T. S., eds, Proceedings of Machine Learning Research Vol. R3, Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics. PMLR, pp 278–285, `https://proceedings.mlr.press/r3/spirtes01a.html`
Spirtes P., Meek C., Richardson T., 1995, in Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. pp 499–506
Spirtes P., Glymour C., Scheines R., 2001, Causation, Prediction, and Search. The MIT Press, doi:10.7551/mitpress/1754.001.0001, `https://doi.org/10.7551/mitpress/1754.001.0001`
Spirtes P. L., Meek C., Richardson T. S., 2013, arXiv e-prints, p. arXiv:1302.4983
Stern S. A., 2002, AJ, 124, 2297
Thébault P., Doressoundiram A., 2003, Icarus, 162, 27
Wong I., Brown M. E., 2016, AJ, 152, 90
Wong I., Brown M. E., 2017, AJ, 153, 145
Zhang J., 2008, Artificial Intelligence, 172, 1873
Zhang K., Peters J., Janzing D., Schölkopf B., 2011, in Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence. pp 804–813
Zhang K., Peters J., Janzing D., Schoelkopf B., 2012, arXiv e-prints, p. arXiv:1202.3775
Zheng Y., et al., 2024, Journal of Machine Learning Research, 25, 1
von Zeipel H., 1910, Astronomische Nachrichten, 183, 345

This paper has been typeset from a TEX/LATEX file prepared by the author.