Monthly Notices
of the Royal Astronomical Society

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

Journal:	*Monthly Notices of the Royal Astronomical Society*
Manuscript ID	Draft
Manuscript type:	Letter
Date Submitted by the Author:	n/a
Complete List of Authors:	Davis, Benjamin; New York University - Abu Dhabi Campus, Center for Astro, Particle, and Planetary Physics Ali-Dib, Mohamad; NYU Abu Dhabi, CAP3 Zheng, Yujia; Carnegie Mellon University Jin, Zehao; New York University Abu Dhabi, Physics Kun.Zhang@mbzuai.ac.ae, Kun; Carnegie Mellon University; Mohamed bin Zayed University of Artificial Intelligence Macciò, Andrea; New York University - Abu Dhabi, ; Max-Planck-Institut fur Astronomie,
Keywords:	Kuiper belt: general < Planetary Systems, methods: statistical < Astronomical instrumentation, methods, and techniques

MNRAS 000, 1–7 (2025) Preprint 15 June 2025 Compiled using MNRAS LATEX style file v3.3

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 have revealed correlations between the eccentricity and inclination of TNOs and their colours. This has rekindled the long-standing debate on whether these colours reflect the conditions of TNO formation or their subsequent collisional evolution. In this study, we address this question with 98.7% certainty, using a model-agnostic, data-driven approach based on causal graphs. First, as a sanity check, we demonstrate how our model can replicate the currently accepted paradigms of TNOs' dynamical history, blindly and without any orbital modelling or physics-based assumptions. In fact, our causal model (without knowledge of Neptune) predicts the existence of an unknown perturbing body, i.e., Neptune. We then show how this model predicts, with high certainty, 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. Moreover, our causal model excludes formation scenarios that invoke substantial colour modification by subsequent irradiation. We therefore 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 often difficult to decipher. 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).

Although the history of these studies is long, here we focus on Marsset et al. (2019) who found a strong correlation between the inclination and colours of TNOs. More precisely, using the Colours of the Outer Solar System Origin Survey (COl-OSSOS; Schwamb et al. 2019) observations, they concluded that Very Red Objects (VROs) have a cut-off maximum inclination of around 21°, in contrast to the more gray Less Red Objects (LROs). These results were expanded by Ali-Dib et al. (2021), who found an analogous trend where the eccentricity of VROs is cut-off at 0.42. They 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. In this scenario, using causality theory jargon, eccentricity (e) and inclination (i) are said to be caused by the colours, which is indicative of the formation location.

The primordial origin hypothesis of the TNO colour diversity argues that TNO colours reflect compositional gradients in the protoplanetary disk, preserved since formation (Nesvorný et al. 2020; Ali-Dib et al. 2021). 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.

However, in a alternative way, many works (Luu & Jewitt 1996; Stern 2002) 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 i) 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.

A third possibility proposed that initially diverse bulk compositions undergo selective volatile evaporation post-formation, estab-

⋆ E-mail: ben.davis@nyu.edu (BLD)

lishing steep compositional gradients across the primordial disk that, coupled with subsequent UV photolysis and particle irradiation, yield distinct surface chemistries (Brown et al. 2011; Wong & Brown 2017). A key difference between this and the 'primordial origin' hypothesis is the necessity of post-formation irradiation. From a causality lens, this introduces a causal relationship between the current semimajor axis and the colour of TNOs.

In this paper, we use a purely data-driven, model-agnostic, statistical causal discovery method to study the relationships between the dynamical parameters of TNOs, and between those and the TNO colors. 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 detail an overview of our causal discovery methods (§2) including a description of our data sample (§2.1), present the results of our analysis (§3), and conclude with a discussion and overall summary (§5). Additionally, we include an Appendix (§A) to cover the details of our ancillary test with Gaussianisation.

2 METHODOLOGY

2.1 Data

Our dataset is based on (but not exclusively) the Col-OSSOS survey (Schwamb et al. 2019). It was taken from Marsset et al. (2019) and Ali-Dib et al. (2021). It consists of a total of 229 TNOs including hot classicals, centaurs, and resonant/scattered objects, in a dataset for which discovery biases were modelled. For each TNO, we have three orbital elements: semimajor axis (a), eccentricity (e), and inclination (i); 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. See Fig. 1 for a pairplot showing all the pairwise relations between our data.[1] Additionally, Fig. 1 shows the subpopulation in our data by separating each TNO by its classifications as either a Classical (48), Resonant (102), Centaur (36), Scattered (28), or Detached (15) object.

Our dataset is further summarized in Fig. 2. We define VROs as TNOs with spectral slopes greater than 20.6%/(10^3 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 Causal discovery: identifying cause-effect relationships

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 randomized 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 deciphering gene regulatory networks (Sachs et al. 2005) to mapping intricate protein signalling pathways (Friedman 2004). 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). 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. (Spirtes et al. 2001) and Pearl (Pearl 2009).

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 dependency relations. By analysing conditional independence 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, revealing which variables *cause* others under appropriate conditions. Ideally, the output is a Directed Acyclic Graph (DAG) for a unique solution or a Completed Partially Directed Acyclic Graph (CPDAG) for a Markov equivalence class. However, when some variables remain unmeasured, certain causal relationships may be undetermined, leading to a Partial Ancestral Graph (PAG). For further reading on causal discovery and causality, see *Causation, Prediction, and Search* (Spirtes et al. 2001), *Causality* (Pearl 2009), or the review in Jin et al. (2025b, §2).

2.3 Causal structures with latent variables

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 orbits, we use the FCI implementation in the `Python` package `causal-learn` (Zheng et al. 2024) to infer the underlying causal structure.

FCI discovers causal relationships by performing a series of conditional independence (CI) tests. These tests examine whether the statistical dependence between two variables disappears when controlling for other variables. If two variables become independent when conditioning on a third, this suggests that the third variable may be an intermediary or a common cause.

[1] See §A for a further test on applying non-linear transformations to Gaussianise our data.

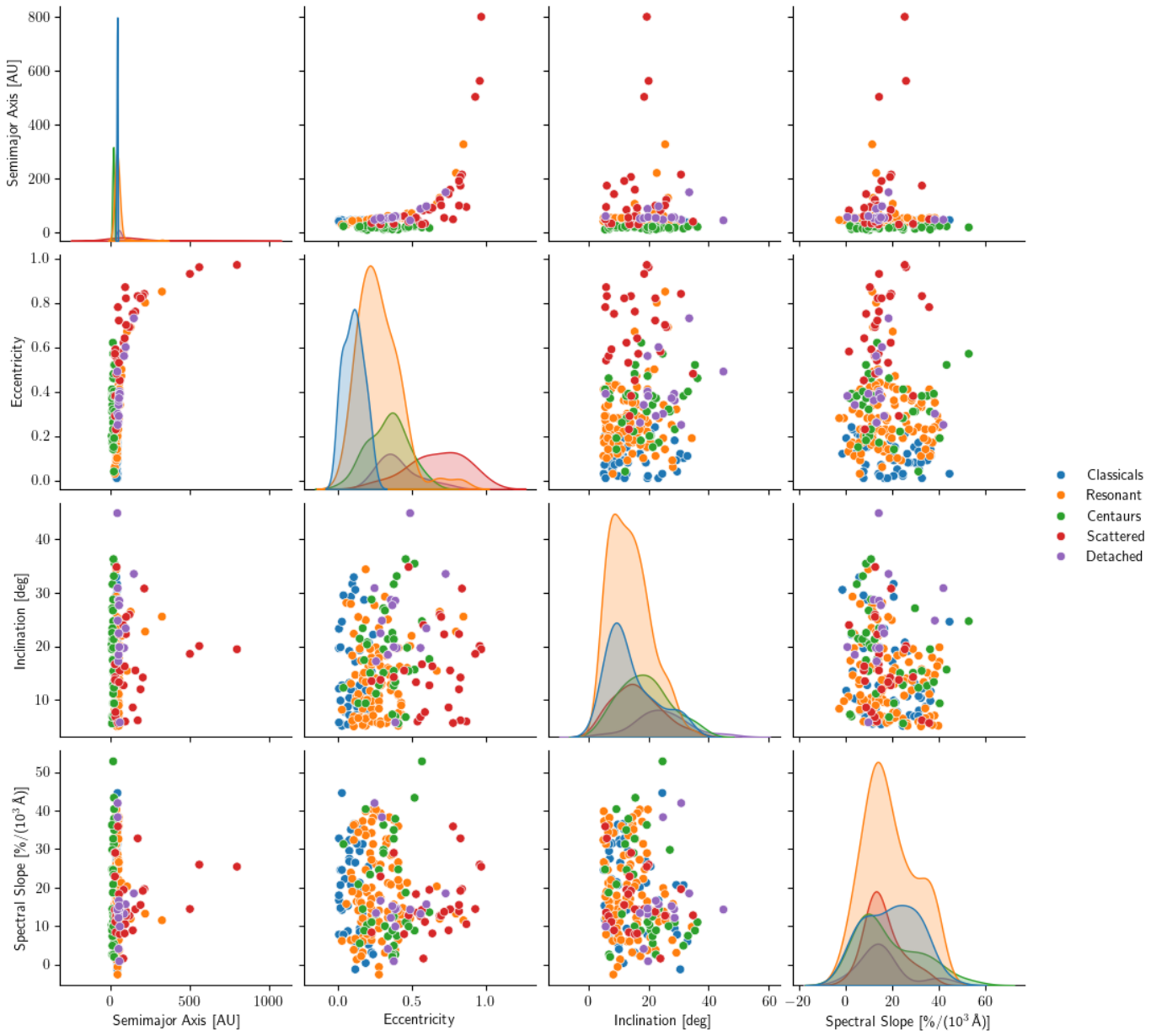

Figure 1. Pairplot of all 229 TNOs in our study. The TNOs are further divided into their individual populations: 48 Classicals (•), 102 Resonant (•), 36 Centaurs (•), 28 Scattered (•), and 15 Detached (•).

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 PAG, which provides more nuanced causal information. The edges in a PAG have different interpretations:

- $X \longrightarrow Y$: X is a *cause* of Y.
- $X \circ\!\!\longrightarrow Y$: Y is not an *ancestor* of X. Intuitively, this implies Y cannot be a cause of X, whether directly or indirectly.
- $X \circ\!\!-\!\!\circ Y$: No set d-separates X and Y. In other words, they may be causally adjacent or share a latent common cause.
- $X \longleftrightarrow Y$: There is a latent common cause of X and Y.

Therefore, by accounting for latent variables in the discovery process, we can 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 RESULTS

3.1 Data-driven results

Our primary findings are summarized in Fig. 3 showing the statistically most likely PAG fitting our data, at 98.7% confidence. This main result utilizes the FCI algorithm (Spirtes 2001; Spirtes et al. 2013;

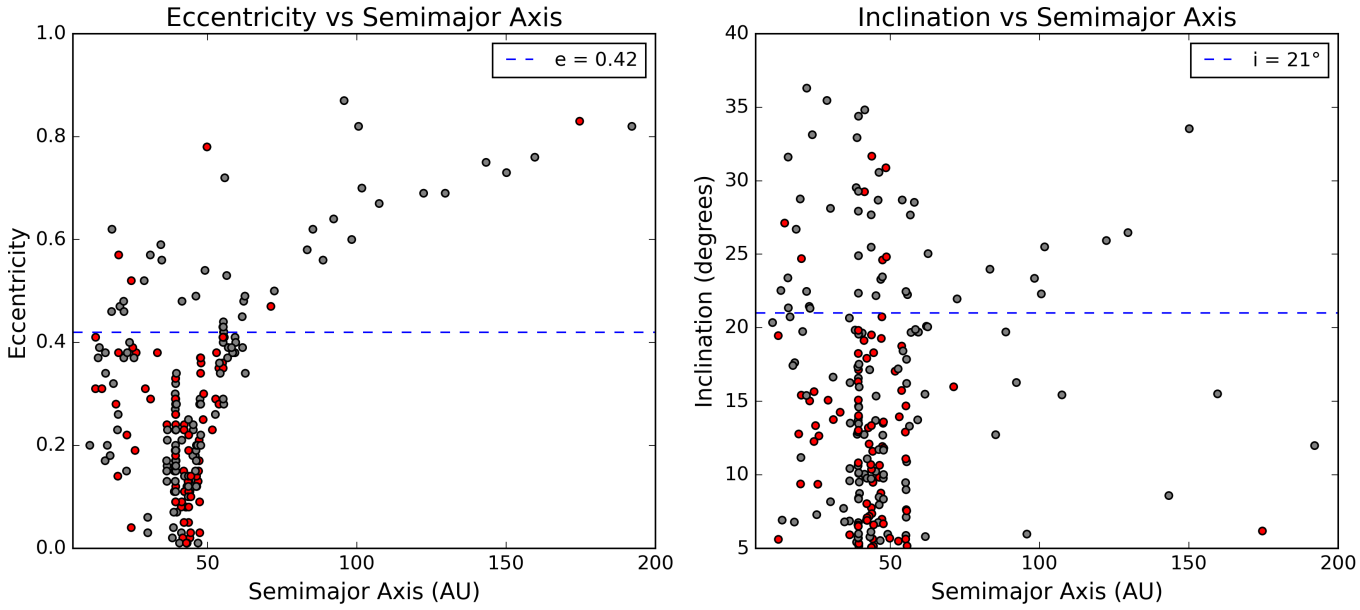

Figure 2. The Marsset et al. (2019) and Ali-Dib et al. (2021) sample shown as a–e (left) and a–i (right) plots. Colours were defined such that red (●) is for Very Red Objects (spectral slopes higher than 20.6%/(10^3 Å)) and gray (●) is for Less Red Objects. The plot clearly shows the paucity of VROs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

Zheng et al. 2024), with linear Fisher-Z conditional independence tests (Fisher 1921), and the threshold for each conditional independence test is $\alpha = 0.013$ (i.e., all tests must pass at the 98.7% level) on transformed data via Gaussianisation. The motivation and details of the Gaussian transformation can be found in §A. We still get the same PAG using a linear Fisher-Z test without any transformation for $\alpha = 0.02$. It is also possible to directly use a non-linear conditional independence test. Here, we adopt a Kernel-based conditional independence (KCI) test (Zhang et al. 2012), with a polynomial kernel and reproduce the same PAG as in Fig. 3 at $\alpha = 0.09$.[2]

We emphasize that this PAG was obtained with a purely data-driven approach, without astrophysical insights. Moreover, we consistently reproduce the same PAG as in Fig. 3 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 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.

3.2 Step-by-step derivation of the PAG

Here we provide a manual derivation of the resulting PAG for readers new to causal discovery and the FCI algorithm. The PAG is built on the list of conditional independencies shown in Table 1. Starting from

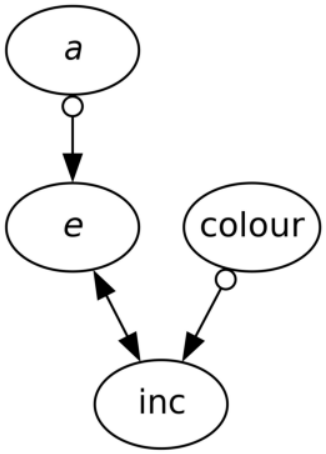

Figure 3. Partial Ancestral Graph (PAG) for 229 TNOs, calculated with the Fast Causal Inference (FCI) algorithm (Spirtes 2001; Spirtes et al. 2013; Zheng et al. 2024), for linear Fisher-Z conditional independence tests (Fisher 1921) on transformed data, with $\alpha = 0.013$ (significance level of individual partial correlation tests). On untransformed data, we recover the identical PAG with linear Fisher-Z tests and $\alpha = 0.02$, while the same PAG is produced with $\alpha = 0.09$ when we run Kernel-based conditional independence (KCI) tests (Zhang et al. 2012), with a polynomial kernel. 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.

the hypothesised, undirected, and fully-connected graph in Fig. 4 (panel 1) among our four nodes (a, e, inc, and colour), we can remove edges between independent nodes. For example, given $a \perp\!\!\!\perp$ inc, it is then not possible to have a cause inc ($a \rightarrow$ inc), inc cause a (inc $\rightarrow a$), nor a third latent variable (L) cause both a and inc

[2] We did not apply non-linear tests in the first place because a non-linear method is prone to overfitting for the relatively small size of our data (229 TNOs).

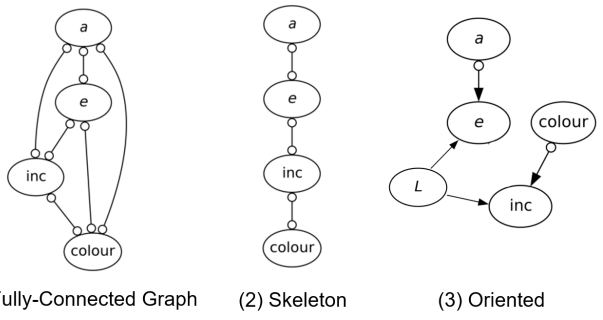

Figure 4. The visualisation of the FCI algorithm. Starting with a fully-connected graph, edges are removed and finally oriented according to the conditional independencies.

($a \leftarrow L \rightarrow$ inc, shortened as $a \leftrightarrow$ inc). Therefore, the edge between a and inc ($a \circ\!\!-\!\!\circ$ inc) can be removed.

Similarly, $a \circ\!\!-\!\!\circ$ colour and $e \circ\!\!-\!\!\circ$ 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.[3]

Now, we are left with with a skeleton graph as Fig. 4 (panel 2), and we shall orient the remaining edges according to conditional independencies. Given the current skeleton, $a \perp\!\!\!\perp$ colour, $a \perp\!\!\!\perp$ inc, and $e \perp\!\!\!\perp$ colour 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., $a \rightarrow e \rightarrow$ inc or $a \leftarrow e \leftarrow$ inc) and a fork structure (i.e., $a \leftarrow e \rightarrow$ 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., $a \circ\!\!\rightarrow e \leftarrow\!\!\circ$inc). Similarly, we can find $e \circ\!\!\rightarrow$ inc $\leftarrow\!\!\circ$colour according to $e \perp\!\!\!\perp$ colour. The need for both $e \leftarrow\!\!\circ$inc and $e \circ\!\!\rightarrow$ inc calls for a latent confounder L causing both e and inc (i.e., $e \leftarrow L \rightarrow$ inc, or $e \leftrightarrow$ inc in a more compact notation). We therefore arrive at the final PAG in Fig. 3 and Fig. 4 (panel 3).

3.3 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:

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

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.

[3] $e \not\perp\!\!\!\perp$ inc does not guarantee an edge between e and inc, as $e \leftarrow a \rightarrow$ inc or $e \leftarrow$ colour $\rightarrow$ 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.

4 ASTROPHYSICAL INTERPRETATION

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:

- 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.

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 (von Zeipel 1910; Lidov 1962; Kozai 1962) anti-correlated oscillations between these two quantities (both inside and outside of MMRs), that plays a central role in the dynamics of TNOs. Here, $e \leftrightarrow i$ 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 discovered in 1846, our result here would strongly suggest the presence of an unknown perturbing body. Together, the first two links successfully reestablish the main dynamical processes shaping the Kuiper belt (scattering, MMRs, and von Zeipel-Lidov-Kozai oscillations) without any physical inputs.

Finally, **the third piece** of the puzzle is the connection colour $\circ\!\!\rightarrow i$ *ruling out* the possibility of colour $\leftarrow i$. The 'colour' (i.e., a proxy for the formation location in our null hypothesis) is hence causing the inclination. 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). Note that this link, however, leaves open the possibility of an unobserved confounder causing both colour and the inclination. 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 $\leftarrow i$ is not allowed, rules out the model of Luu & Jewitt (1996) and Stern (2002), where collisional evolution shapes the colors of TNOs. Moreover, our result that colour $\leftarrow a$ is not allowed either, rules out the model of Brown et al. (2011) and Wong & Brown (2017), where a would control the amount of irradiation a TNO is subjected to. We are hence left only with the 'primordial origins' model where **the colour is set entirely by the chemical composition of the formation location**.

Some further interesting features are found in the PAG:

- **The lack of correlation between the colour and semimajor**

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. For example, '$X \perp\!\!\!\perp Y \mid Z$' means '$X$ is independent to Y when conditioned on Z.' Three sets of conditional independence tests are performed: Fisher-Z test on transformed data, Fisher-Z test on untransformed data, and KCI test on untransformed data. The p-value for the null hypothesis for each test is shown. A p-value closer to 0 suggests dependence, and a p-value closer to 1 favours independence. As discussed in §3.1, we choose the threshold for each conditional independence test α to be 0.013, 0.02, and 0.09 for Fisher-Z test on transformed data, Fisher-Z test on untransformed data, and KCI test on untransformed data, respectively. The PAG shown in Fig. 3 directly comes from this list of conditional independencies.

Conditional independencies	Conditional independence test p-value		
	transformed, Fisher-Z	untransformed, Fisher-Z	untransformed, KCI
$a \not\perp\!\!\!\perp e$	0.000	0.000	0.000
$a \perp\!\!\!\perp$ inc	0.750	0.144	0.421
$a \perp\!\!\!\perp$ colour	0.958	0.918	0.488
$e \not\perp\!\!\!\perp$ inc	0.005	0.005	0.085
$e \perp\!\!\!\perp$ colour	0.092	0.211	0.135
inc $\not\perp\!\!\!\perp$ colour	0.001	0.001	0.009
$a \not\perp\!\!\!\perp e \mid$ inc	0.000	0.000	0.068
$e \not\perp\!\!\!\perp$ inc $\mid a$	0.004	0.018	0.070
$e \not\perp\!\!\!\perp$ inc $\mid$ colour	0.012	0.010	0.041
inc $\not\perp\!\!\!\perp$ colour $\mid e$	0.002	0.003	0.004

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 Figs. 11 and 12 of Ali-Dib et al. (2021). See also Fig. 3 of Nesvorný et al. (2016).

- **The indirect causation between the colour (initial location) and the eccentricity through 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. In all cases, we note that the correlation coefficient between the colour and eccentricity is around 0.1, allowing for a minor correlation between the two that can be seen in less probable PAGs.

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, with 98.7% certainty that *TNO colour is causally antecedent to inclination, not a consequence of it.* 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. This will be explored further in the future.

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 is 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, this work is a proof of principle for the use of causality models in planetary sciences. In our case, it even 'rediscovers' the existence of Neptune. With large datasets ranging from asteroids to exoplanets, many discoveries await.

ACKNOWLEDGEMENTS

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
Brown M. E., Schaller E. L., Fraser W. C., 2011, ApJ, 739, L60
Colombo D., Maathuis M. H., Kalisch M., Richardson T. S., 2012, The Annals of Statistics, pp 294–321
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
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
Kozai Y., 1962, AJ, 67, 591

Lidov M. L., 1962, Planet. Space Sci., 9, 719
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
Runge J., et al., 2019, Nature communications, 10, 1
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
Wong I., Brown M. E., 2017, AJ, 153, 145
Yeo I., Johnson R. A., 2000, Biometrika, 87, 954
Zhang J., 2008, Artificial Intelligence, 172, 1873
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

APPENDIX A: DATA PREPROCESSING

Since the Fisher Z-test used in our causal discovery algorithm assumes linear Gaussian distributions of the model, we apply deterministic, variable-wise transformations to our data for preprocessing. Specifically, we employ the Yeo-Johnson transformations (Yeo & Johnson 2000) as our primary preprocessing. For the semimajor axis (a) and eccentricity (e), we employ a combination of Yeo-Johnson and tanh-type transformations to better handle their non-linear relationships. For inclination and spectral slope, the standard Yeo-Johnson Gaussianisation is sufficient. As these transformations act independently on each variable and are deterministic, they preserve the underlying causal structure.

The preprocessed data for all 229 TNOs is displayed as a pairplot in Fig. A1. Using preprocessed data, we reproduce the same PAG structure as shown in Fig. 3 with $\alpha = 0.013$. Visual inspection of the transformed data confirms the effect of the transformation, making the data more suitable for the Fisher Z-test.

In fact, the transformed scatter plots also provide insights into causal directions. For instance, when examining the relationship between spectral slope and inclination in the direction of spectral slope causing inclination (i.e., a linear model of $i = k \cdot \text{slope} + \epsilon$), the noise term ϵ appears more independent compared to the reverse direction, supporting this causal orientation in our final PAG.

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

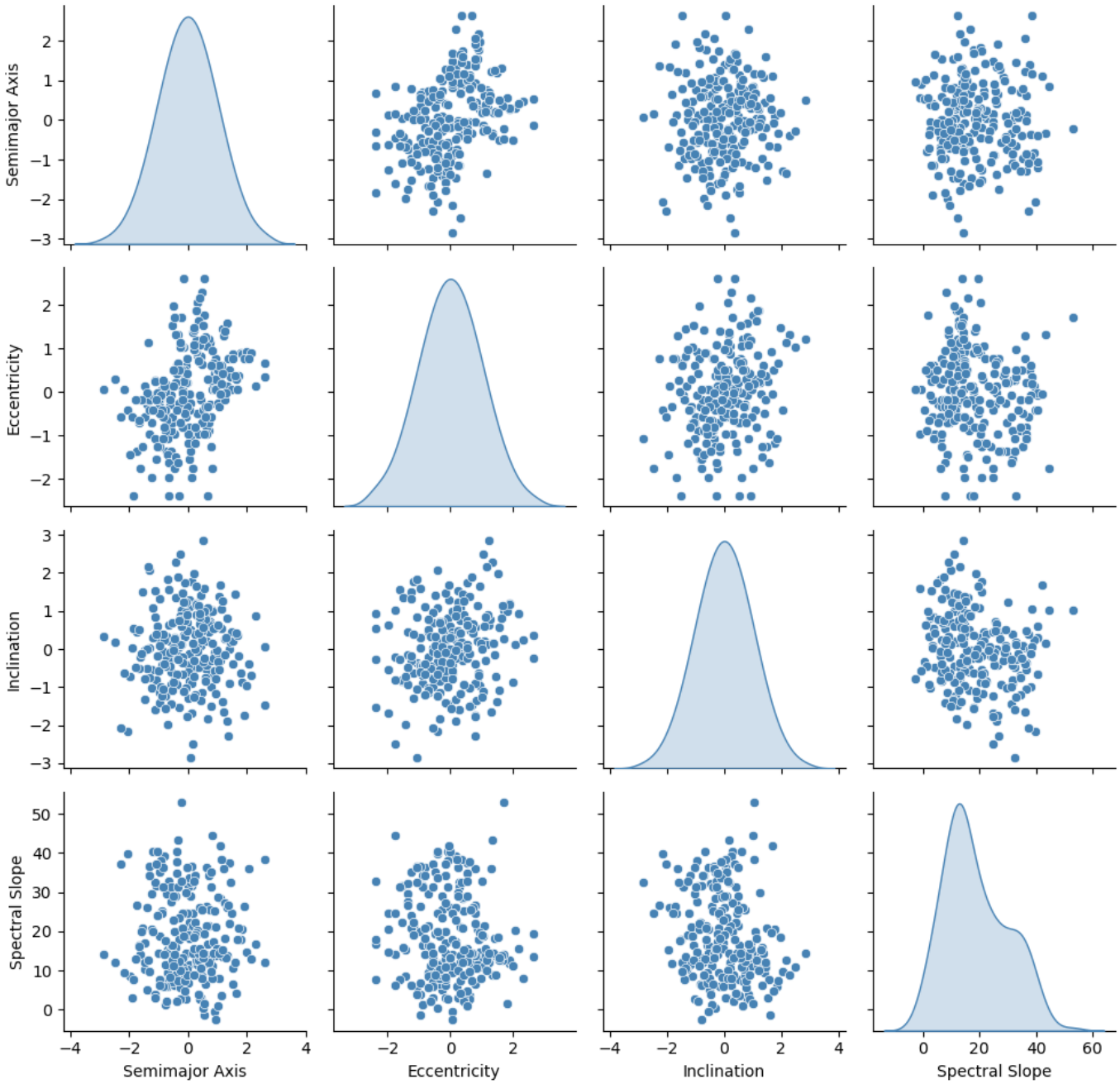

Figure A1. Pairplot of all 229 TNOs in our study that is similar to Fig. 1, except we have performed deterministic, variable-wise transformations to preprocess the data for the Fisher-Z test.