



OPEN ACCESS

EDITED BY

Tao Zeng,
Guangzhou Laboratory, China

REVIEWED BY

Dola Sundeep,
Indian Institute of Information Technology
Design and Manufacturing, India
Michiel Stock,
Ghent University, Belgium

*CORRESPONDENCE

Hans A. Kestler,
✉ hans.kestler@uni-ulm.de

[†]These authors have contributed equally
to this work

RECEIVED 14 July 2025

ACCEPTED 13 October 2025

PUBLISHED 03 November 2025

CITATION

Stolnicu A, Eckhardt-Bellmann P, Kestler AMR
and Kestler HA (2025) Identification of ordinal
relations and alternative suborders within
high-dimensional molecular data.
Front. Bioinform. 5:1665892.
doi: 10.3389/fbinf.2025.1665892

COPYRIGHT

© 2025 Stolnicu, Eckhardt-Bellmann, Kestler
and Kestler. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Identification of ordinal relations and alternative suborders within high-dimensional molecular data

Ana Stolnicu^{1†}, Peter Eckhardt-Bellmann^{1†},
Angelika M. R. Kestler² and Hans A. Kestler^{1,3*}

¹Institute of Medical Systems Biology, Ulm University, Ulm, Germany, ²Department of Internal Medicine I, Ulm University Hospital, Ulm, Germany, ³Leibniz Institute on Aging – Fritz Lipmann Institute, Jena, Germany

Introduction: Numerous biological systems exhibit ordinal connections between categories. Developmental and time-series information inherently depict sequences like “early,” “intermediate,” and “late” phases, showing that these specific processes follow a progression. Ordinal classification techniques are often applied in biological and medical contexts, ranging from the evaluation of pain intensity, to the detection of evolving diseases, such as cancer. These ranking systems may assist clinicians in establishing diagnoses and developing tailored treatment plans. For instance, tumor staging might guide early detection strategies and targeted therapies, improving patient outcomes. However, applying ordinal classification to biological data presents considerable challenges. In addition to their high dimensionality, these datasets can be highly heterogeneous, often reflecting branching processes that occur simultaneously during progression. Factors such as intratumoral diversity, asynchronous progress, and context-specific signaling activity may interfere with the identification of such alternative development routes.

Methods: To address these challenges, we propose a framework for uncovering ordinal relationships within molecular data. Specifically, directed threshold classifiers are introduced as base learners for ordinal classifier cascades, enabling the detection of both total and partial orderings between molecular states.

Results: This approach preserves the inherent ordinal structure by projecting high-dimensional data onto one single dimension while simultaneously decreasing complexity. Additionally, the distinct features of the resulting thresholds allow the prediction of potential alternative paths among the suborders.

KEYWORDS

alternative progression patterns, classifier cascades, directed threshold classifiers, ordinal classification, high-throughput data

1 Introduction

Various physiological processes and health conditions naturally follow an ordinal arrangement, in which stages progress hierarchically (Prigogine and Nicolis, 1971). Organizing disease phases into meaningful semantic groups can be a valuable predictive tool in clinical practice (Lee et al., 2004). In oncology, tumor classification aids in prognosis and treatment strategies, guiding the choice of interventions and targeted therapies

based on predicted tumor behavior (Beadsmoore and Screaton, 2003; Forner et al., 2014; Cortés et al., 2014). Similarly, categorizing the stages of neurodegenerative disorders, such as Alzheimer's disease (Sperling et al., 2011; Davis et al., 2018; Tahami Monfared et al., 2022) might facilitate prompt therapeutic decisions and patient monitoring (Scharre, 2019). Pain classification adheres to comparable principles, where pain intensity reported by patients can be arranged into numerical or categorical scales (Hadjistavropoulos and Craig, 2002; Haefeli and Elfering, 2006). These may also be used in anesthesiology and pain management to guide treatment suitability (Breivik et al., 2008). Despite its clinical utility, ordinal classification in biological and medical data presents considerable computational challenges. Molecular datasets, such as gene expression profiles, are not only high-dimensional, consisting of thousands of interrelated features, but also may encode multiple, potentially parallel biological processes, each with its own progression dynamics (Brody, 2009; Wu et al., 2019; Yang et al., 2022; Gerlinger et al., 2012). Additionally, high-throughput data often suffer from noise introduced by experimental variability, batch effects, and underlying biological diversity, which can hinder ordinal relationships and complicate model training (Tu et al., 2002; Goh et al., 2017). Moreover, numerous biological processes lack clear stage transitions, exhibiting overlapped molecular signatures and divergent trajectories, leading to ambiguous classification boundaries and partial ordering of states (Seoane and De Mattos-Arruda, 2014). Further suggesting that these mechanisms could potentially evolve through various parallel pathways (Olschwang et al., 1997; Traverso et al., 2002).

The proposed architecture is tailored for the detection of ordinal structures within one-dimensional data, derived from high-throughput datasets. The categories, i.e., classes, are delineated by thresholds that partition the input space into distinct intervals. An essential aspect of this method is the ability to recreate potential alternative ordinal trajectories from the resulting suborders. This can be accomplished based on the properties of the decision boundaries. As a result, the model can provide a distinct benefit in scenarios in which the ordinal structure may not be strictly predefined, allowing for more nuanced and adaptable classification decisions. However, note that our introduced architecture is designed for the detection of ordinal structures and (parallel) substructures, not for the classification process at hand.

2 Related work

Ordinal classification is a type of supervised learning in which the classes exhibit an intrinsic order that does not necessarily adhere to specific numerical intervals (Frank and Hall, 2001). In contrast to conventional ordinal classification approaches, which typically assume a fixed class order and often fail to capture the optimal ordinal correlations between classes, ordinal classifier cascades (OCCs) (Lattke et al., 2015) decompose the task into a series of simplified binary classification problems. In this framework, a cascade of classifiers is used, where each classifier determines whether a given instance belongs to a specific category or a higher-ranked one. The cascade approach evaluates samples sequentially, attributing a label according to the first classifier that provides a confident prediction. This structure not only

streamlines the classification problem at each stage, but also allows the exploration of potential class sequences. In this context, the CASCADES algorithm (Lausser et al., 2019) extends the sequential framework by improving its efficiency, replacing the exhaustive search with exploratory screening of candidate orders. To handle the computational complexity of this search, it employs early rejection criteria based on class-wise sensitivity limits, discarding underperforming cascades prior to complete training. Additionally, binary classifiers in the cascade are trained to distinguish between a class and its successor, enabling pairwise trained classifiers to be stored and reused for different input orders, thus decreasing runtime and minimizing redundant computations. Because the algorithm is independent of the classifier type, allowing the integration of any suitable binary training method, this approach enhances both efficiency and flexibility. Finally, it produces a set of candidate cascades that satisfy the established performance criteria, which can be further evaluated for ensemble integration or downstream model selection.

Formally, in the context of ordinal classification, we are given a set of $\mathcal{N}$ samples, $\mathcal{D} = \{(x_k, y_k)\}_{k=1}^{\mathcal{N}}$, where $x_k \in \mathcal{X}$ denotes the feature vector of the k -th sample and $y_k \in \mathcal{L}$ indicates its associated label. Here, $\mathcal{X} \subseteq \mathbb{R}^d$ is the feature space and $\mathcal{L} = \{l_1, l_2, \dots, l_{|\mathcal{L}|}\}$ corresponds to the finite set of class labels. The objective is to predict the label for each sample k taking into account its feature vector. Thus, a binary classifier, $c_{(i,i+1)}$, of an OCCs ensemble, ε_C , is trained to differentiate between samples belonging to adjacent classes, l_i and l_{i+1} , in the given semantic order, $l_1 < l_2 < \dots < l_{|\mathcal{L}|}$, as:

$$\varepsilon_C = \{c_{(i,i+1)}: \mathcal{X} \mapsto \{l_i, l_{i+1}\} \mid i = 1, \dots, |\mathcal{L}| - 1\}. \quad (1)$$

The index i designates the position of the classes in the given order. Throughout the classification procedure, every x_k is evaluated by the sequence of classifiers arranged according to the order under investigation, that is, if the input order is $o = l_1 < l_2 < \dots < l_{|\mathcal{L}|}$, the classifiers are organized as $\{c_{(1,2)}, c_{(2,3)}, \dots, c_{(|\mathcal{L}|-1, |\mathcal{L}|)}\}$. For a sample k , if a classifier $c_{(i,i+1)}(x_k)$ generates a positive prediction for the first label, l_i , the corresponding label will be assigned to x_k , and the cascade ends. Otherwise, the sample is passed to the next classifier in the sequence, continuing the process until the final classifier $c_{(|\mathcal{L}|-1, |\mathcal{L}|)}$ is reached, in which case, if the second label is predicted, then the predicted label y'_k is equal to $l_{|\mathcal{L}|}$, as defined in Equation 2:

$$y'_k = \begin{cases} l_j, & \text{where } j = \min \{i \in \{1, \dots, |\mathcal{L}| - 1\} \mid c_{(i,i+1)}(x_k) = l_i\}, \\ l_{|\mathcal{L}|}, & \text{if } c_{(i,i+1)}(x_k) = l_{i+1} \quad \forall i < |\mathcal{L}|. \end{cases} \quad (2)$$

In order to guide the selection of the most effective cascades, the class-wise sensitivity serves as primary efficiency criterion for the classifiers. An example of an OCC architecture is depicted in Figure 1.

Bellmann and Schwenker (2020) proposed another approach for the detection of ordinal class structures, in which it is not necessary to explicitly evaluate all possible class orderings. The idea is to determine the performance (resubstitution accuracy) of linear Support Vector Machines (SVMs) (Vapnik, 2000) for each class pair, i.e., $|\mathcal{L}| \cdot (|\mathcal{L}| - 1)/2$ binary subtasks. The resulting performance values, $a_{i,j}$, imply how well the classes, $\{l_i, l_j\}$, can be separated from each other. As the next step, the values $a_{i,j}$ are combined into $|\mathcal{L}|$

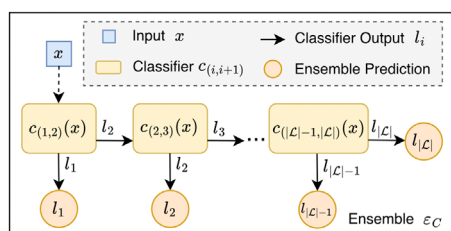


FIGURE 1

Ordinal classifier cascade (OCC) ensemble. The OCC architecture consists of $|L| - 1$ binary classifiers $c_{(i,i+1)}$ that can either predict label l_i or label l_{i+1} . If the *greater* class (l_{i+1}) is predicted, the input is passed to the next classifier in the sequence. Otherwise, if the *lower* class (l_i) is predicted by $c_{(i,i+1)}$, then this output is taken as the ensemble's final prediction for input x . The last classifier in the sequence, $c_{(|L|-1,|L|)}$, cannot further pass input x , and therefore, once reached, always provides the ensemble's final output, by predicting either $l_{|L|-1}$ or $l_{|L|}$. An OCC ensemble is defined by its set of classifiers, as in Equation 1.

symmetric matrices $\mathcal{A}$, $\mathcal{A} = (a_{ij})_{i,j=1}^{|L|}$, with different arrangements of the row (and column) elements. While the symmetry of each $\mathcal{A}$ is obtained by definition, due to $a_{ij} = a_{ji}$, for all $i \neq j$, the authors defined $a_{ii} := 0$, $\forall i = 1, \dots, |L|$. An ordinal class structure is found if and only if there exist exactly two matrices $\mathcal{A}$ for which the row (and column) entries are monotonously decreasing towards the diagonal elements. From the symmetry characteristic, it follows that each ordinal structure is found together with its reverse order.

Bellmann and Schwenker (2020) further extended their work in (Bellmann et al., 2022). They generalized their working definition of ordinal classification tasks by introducing a theoretical framework which makes it possible to detect ordinal class structures without utilizing any classification model. As an example, they proposed using a multidimensional adaptation of Fisher's discriminant ratio (Fisher, 1936). Using their framework, they proved that, in general, 3-class classification problems can be regarded as ordinal classification tasks consisting of two edge classes and a class identified as the central one. Note that the authors reduced the detection complexity from evaluating all possible class orderings, $|L|!$ evaluations, to only $|L| \cdot (|L| - 1)/2$. However, they did not discuss the potential for detecting substructures, a useful property that was elaborated by Lausser et al. (2020) based on the CASCADES algorithm. In contrast to the methods discussed above, in our current approach, the mining for ordinal suborders is not conducted in the provided, and often high-dimensional, feature space, but in combination with the one-dimensional real space. Moreover, with our approach presented in this work, we are able to identify alternate progressions.

3 Materials and methods

3.1 Directed threshold classifiers

The purpose of the Directed Threshold Classifiers (DTCs) introduced in this work is to recognize ordinal relations within univariate data $\mathcal{X} \subseteq \mathbb{R}$. A DTC $f_\tau: \mathcal{X} \rightarrow \{l_i, l_j\}$, defined by a

threshold $\tau \in \mathbb{R}$, is built to differentiate between two distinct categories l_i, l_j :

$$f_\tau(x) = \begin{cases} l_j, & \text{if } x \geq \tau, \\ l_i, & \text{otherwise.} \end{cases} \quad (3)$$

The threshold τ divides the input space into two decision areas, in which all elements belonging to class l_j , which have values greater than τ , are assigned on the right side, whereas instances of class l_i , with values below τ fall into the region on the left side, as shown in Equation 3. A set of DTCs can be organized sequentially according to a specified input order $o = l_1 < \dots < l_{|L|}$ to be further applied as base classifiers within the OCCs framework. The samples being examined are assumed to be arranged along a one-dimensional axis, and the thresholds, corresponding to specific points, are constrained to follow a strictly increasing order on the same line, $\tau_1 < \dots < \tau_{|L|-1}$. This guarantees that the decision regions form contiguous segments within the space, leading to a connected and non-overlapping partitioning of the domain that mirrors a consistent progression aligned with the ordinal nature of the targeted labels. Moreover, the non-intersecting characteristic of the regions inherently creates parallel decision boundaries, as each one is orthogonal to the axis of progression. For the computation of the one-dimensional thresholds, we apply linear SVM models, making use of their margin maximization characteristic.

3.2 Data transformation to one dimension

As univariate data rarely appear in real-world scenarios, the first step of the method involves dimension reduction, for which supervised and non-supervised techniques exist. Principal components analysis (PCA) (Kambhatla and Leen, 1997), Linear discriminant analysis (LDA) (Fisher, 1936), t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008), and uniform manifold approximation and projection (UMAP) (McInnes and Healy, 2018) are just a few of the numerous applicable methods that can be used. In this section, we provide a different strategy tailored to meet the specific objective of our study. The process is summed up in the following main steps: From the available category set we select a pair of classes, (l_i, l_j) , to which we apply a linear binary classifier. The data points are then projected onto the orthogonal hyperplane of the resulting linear model. For this binary linear classification, we employed SVMs, in which the data were streamlined to a one-dimensional form by mapping the points using the normal vector.

Note that we prioritized SVM models for the mapping of the high-dimensional data onto one dimension for the following main reasons. First, SVM models are supervised, i.e., classes play an important role during projection. Second, SVMs are deterministic, ensuring reproducibility. In addition, SVM models maximize the margin between the classes of the chosen projection class pair, which we consider to be important when mining for ordinal structures in the one-dimensional space. However, users of our introduced approach can replace the SVM-based projection by any projection of their preferred choice.

Given that the selection of the initial data mapping most likely affects the direction of the DTCs during the overall screening

process, a key aspect to take into account is the choice of this class pair. Despite appearing trivial, it is important to notice that the two classes are maintained apart from each other in the classification process. Consequently, the resulting projection is likely to highlight distinctions between these selected classes, potentially overlooking variations or correlations in the other classes. In the experiments reported in this work, we examined every possible pairwise combination. We observed that using the two least related categories in the developmental process described by the dataset, generally produced the most consistent results.

3.3 Alternative progressions

In the cascaded system, both total orders and potential suborders can be identified. When partial configurations emerge, it may be particularly valuable to investigate whether they reflect alternative advancements of the same underlying progression. In this context, the afore described properties of the thresholds can help uncover and characterize competing developmental paths. For suborders to be considered as potential parallel trajectories of the same process, they must share a subset of thresholds. In the following, we formally define the criteria that determine when a threshold qualifies as shared between suborders. Let $\mathcal{L} = \{l_1, \dots, l_{|\mathcal{L}|}\}$ represent a finite collection of class labels for which no global order is determined. Assume that two suborders, $o \subset \mathcal{L}$ and $o' \subset \mathcal{L}$, can be recognized so that each defines a valid ordinal sequence. Suppose that a classifier system exists according to which the associated decision threshold is identified with minimal class-wise sensitivity, $sens = 1$, for every category within the respective suborders (i.e., all class instances are correctly classified). The threshold sets obtained for the suborders o and o' can be denoted as $\tau_o = \{\tau_1, \dots, \tau_k\}$ and $\tau_{o'} = \{\tau'_1, \dots, \tau'_m\}$, respectively. We are interested in identifying whether a threshold equivalence relation $\tau_i \equiv \tau'_j$, with $\tau_i \in \tau_o$ and $\tau'_j \in \tau_{o'}$, can be established. Two thresholds are deemed equivalent if they induce identical separation boundaries in regions in which two distinct classes have the same adjacent class in their respective suborders. A threshold can be left-shared or right-shared, depending on whether the common neighboring class is on the left or on the right side of the two categories, detailed in Equations 4–10. Formally, given the classes $l_a \in o \setminus o'$, $l_b \in o' \setminus o$ and $l_i \in o \cap o'$, if the subsequent inequalities occur,

$$\mathcal{X}_{l_i} < \tau_i \leq \mathcal{X}_{l_a} \text{ and} \quad (4)$$

$$\mathcal{X}_{l_i} < \tau'_j \leq \mathcal{X}_{l_b}, \quad (5)$$

where $\mathcal{X}_{l_i}$ represents the set of feature values associated with class l_i , then $\tau_i \equiv \tau'_j$. Moreover, a threshold τ_{ls} exists such that $\tau_i, \tau'_j \mapsto \tau_{ls}$, where τ_{ls} represents the left-shared threshold between the respective class transitions. Similarly, if the following inequalities arise,

$$\mathcal{X}_{l_a} < \tau_i \leq \mathcal{X}_{l_i} \text{ and} \quad (6)$$

$$\mathcal{X}_{l_b} < \tau'_j \leq \mathcal{X}_{l_i}, \quad (7)$$

then $\tau_i \equiv \tau'_j$, and a threshold τ_{rs} exists, such that $\tau_i, \tau'_j \mapsto \tau_{rs}$ represents the right-shared threshold of l_a and l_b . It follows that

τ_{ls} will be situated between l_i and the minimum among l_a and l_b , whereas, τ_{rs} has to be greater than the maximum of l_a and l_b and less than l_i :

$$\mathcal{X}_{l_i} < \tau_{ls} \leq \min\{\mathcal{X}_{l_a}, \mathcal{X}_{l_b}\}, \quad (8)$$

$$\max\{\mathcal{X}_{l_a}, \mathcal{X}_{l_b}\} < \tau_{rs} \leq \mathcal{X}_{l_i}. \quad (9)$$

In a wider framework, in which incorrect sample classifications are allowed with a misclassification rate of $\theta = 1 - sens$, with $sens \in [0.5, 1]$, θ can be incorporated to define the thresholds between each pair of adjacent classes l_i and l_{i+1} as:

$$\tau \in [\mathcal{X}_{l_i} + \theta \cdot (\mathcal{X}_{l_{i+1}} - \mathcal{X}_{l_i}), \mathcal{X}_{l_{i+1}} - \theta \cdot (\mathcal{X}_{l_{i+1}} - \mathcal{X}_{l_i})]. \quad (10)$$

Two suborders are required to have a common minimal class sensitivity for each involved class to qualify as viable alternatives, thus left- and right-shared thresholds can be adapted to account for the same amount of misclassifications as outlined below:

$$\begin{aligned} \tau_{ls} &> \mathcal{X}_{l_i} + \theta \cdot \min\{\mathcal{X}_{l_a} - \mathcal{X}_{l_i}, \mathcal{X}_{l_b} - \mathcal{X}_{l_i}\}, \\ \tau_{ls} &\leq \min\{\mathcal{X}_{l_a} - \theta \cdot (\mathcal{X}_{l_a} - \mathcal{X}_{l_i}), \mathcal{X}_{l_b} - \theta \cdot (\mathcal{X}_{l_b} - \mathcal{X}_{l_i})\}, \\ \tau_{rs} &> \max\{\mathcal{X}_{l_a} - \theta \cdot (\mathcal{X}_{l_a} - \mathcal{X}_{l_i}), \mathcal{X}_{l_b} - \theta \cdot (\mathcal{X}_{l_b} - \mathcal{X}_{l_i})\}, \\ \tau_{rs} &\leq \mathcal{X}_{l_i} + \theta \cdot \max\{\mathcal{X}_{l_a} - \mathcal{X}_{l_i}, \mathcal{X}_{l_b} - \mathcal{X}_{l_i}\}. \end{aligned}$$

This ensures that the decision boundaries retain a consistent level of ambiguity across class transitions. The concept of shared thresholds is illustrated in Figure 2.

A visual representation of the designed procedure is provided in Figure 3, beginning with the data projection (A–B), followed by the application of DTCs and the screening procedure to extract ordinal substructures (C–D), and concluding with their aggregation for the retrieval of potential alternative structures (E).

3.4 Reversed orders

Another feature of this approach lies in the implicit retrieval of inverted suborders. More precisely, for a specific class pair (l_i, l_j) , if the sequence $o = l_1 < \dots < l_{|\mathcal{L}|}$ is retrieved, applying the inverse combination, i.e., (l_j, l_i) , for the data transformation yields the reversed order $o' = l_{|\mathcal{L}|} < \dots < l_1$. This behavior arises from the fact that switching the class pairs results in a mapping transformation that mirrors the original structure, thereby naturally producing the converse sequence without additional interventions. By definition, this means that reversed orders are mathematical artifacts. Whether the reversal is biologically meaningful depends on the classification task at hand and has to be discussed for each case individually.

3.5 Analyzed datasets

The method was initially evaluated using synthetic data comprising 10 distinct categories, $l_0, \dots, l_9$, each containing 100 samples described by two features, which were further reduced to one dimension using the technique introduced in Section 3.2.

To validate our approach, we used two publicly available developmental datasets from the Gene Expression Omnibus (GEO)

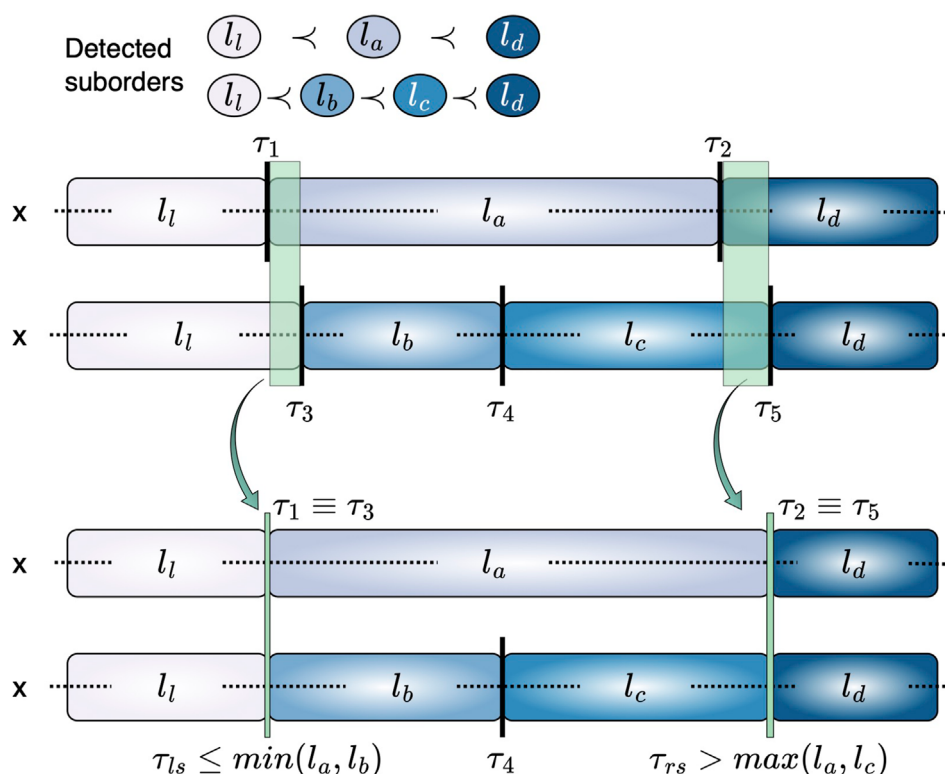


FIGURE 2

Representation of equivalent thresholds across suborders. For suborders $l_l < l_a < l_d$ and $l_l < l_b < l_c < l_d$, τ_1 and τ_3 share l_l on the left, similarly τ_2 and τ_5 share l_d on the right. This characteristic allows to consider τ_1 and τ_3 , as well as τ_2 and τ_5 , as equivalent, enabling the arrangement of the two suborders as alternatives of the same phenomenon.

(Barrett et al., 2012). The expression measurements of 4028 genes of *Drosophila melanogaster* (*D. melanogaster*) (Arbeitman et al., 2002) (included in GEO accession number: GSE4347) were taken at various stages of the fruit fly's life cycle. The developmental phases can be arranged as *embryo* < *larva* < *pupa* < *adult*, with 31, 10, 18 and 8 samples in each category, respectively. The second dataset is composed by pineal glands gene expression profiles collected at five distinct time periods of the zebrafish's (*D. rerio*) maturation process (Toyama et al., 2009) (GEO accession number: GSE13371). They cover three embryonic (3 days, 5 days, and 10 days) and two adult time points (3 months, 1–2 years). The first group consists of 14, 14, and 15 samples, respectively, whereas the second group comprises 12 and 14 samples, respectively.

Furthermore, we used two tumor datasets to test our methodology. The pancreatic ductal adenocarcinoma (PDAC) (Buchholz et al., 2005) which includes 21521 gene expression profiles from human microdissected cells, with 38 samples split into 5 classes: normal ductal cells (6 samples), three intermediate pancreatic intraepithelial neoplasia (PanIN), PanIN-1 (6 samples), PanIN-2 (8 samples) and PanIN-3 (10 samples), as well as the metastatic stage (PDAC) (8 samples). This process is assumed to develop according to the sequence normal < PanIN-1 < PanIN-2 < PanIN-3 < PDAC. The pancreatic neuroendocrine tumors (PanNET) (Sadanandam et al., 2015) (GEO accession number: GSE73514) comprise 35511 mutational profiles from the RIP1 TAG2 mouse model, containing 22 samples organized

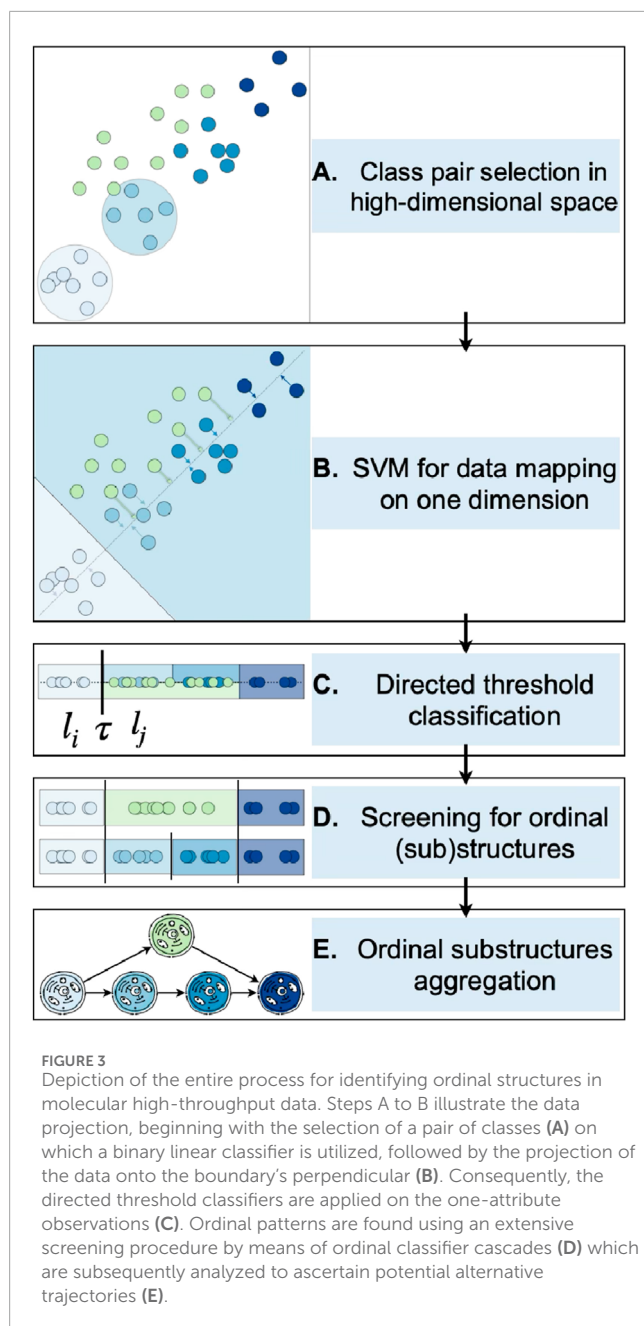
into 6 categories: 3 samples for each normal mature β -cells (NM), hyperplastic islet (HI), angiogenic islet (AI) and liver metastasis (MET), and 5 samples for tumor islet (TI) and met like primary (MLP), each. The assumed progression is NM < HI < AI < TI < MLP < MET.

For all non-synthetic datasets analyzed in this work, we utilized the normalized versions of the samples provided by the original authors to ensure reproducibility. Details of the normalization procedures can be found in the respective dataset publications (Arbeitman et al., 2002; Toyama et al., 2009; Buchholz et al., 2005; Sadanandam et al., 2015). For the zebrafish dataset (Toyama et al., 2009) we additionally applied a $\log_2$ transformation to stabilize variance and diminish asymmetry.

4 Results

4.1 Synthetic data simulations

Upon considering either dimension of the simulated data, no ordinal arrangement encompassing all classes can be discerned with minimal class sensitivity of 1, as illustrated in Figure 4. To investigate the impact of the data projection on the final outcome, we employed all pair combinations of the categories which is the design of the linear decision boundary. The class pairings that returned orders of length six or five are shown in Figure 5. The suborders are illustrated



in a concise graph where overlapping categories, or groups, are shown layered atop each other. For example, in the first graph, the sequence $(l_0 < l_1)$ extends alongside class l_6 , likewise l_2 overlaps with l_7 , l_3 with l_8 , and $(l_4 < l_5)$ with l_9 .

It can be seen that the suborders $(l_0 < l_1 < l_2 < l_3 < l_4 < l_5)$ and $(l_6 < l_7 < l_8 < l_9)$ appear in the outcomes obtained from various combinations. The majority consists of class couples that incorporate categories from the same suborder, for instance l_0 paired with any other class among $\{l_1, \dots, l_5\}$ or l_9 with any class from $\{l_6, l_7, l_8\}$. Six out of 45 combinations, namely (l_0, l_6) , (l_1, l_6) , (l_2, l_7) , (l_3, l_8) , (l_4, l_9) and (l_5, l_9) , produced suborders with lengths less than four and are excluded from the shown results. Particularly poor were the sequences obtained from the data mapping of pair (l_2, l_7) , in which only orders of length two were identified.

4.2 Empirical datasets

We additionally analyzed our approach using the developmental datasets. Alongside the employed projection class pair, Figure 6 presents the outcomes of length four and three achieved for *D. melanogaster* and of lengths from five to three for *D. rerio*. After projecting the data based on class pair (embryo, adult), being the first and last stages in the maturation process, our classification strategy accurately provided the fruit fly's development, $embryo < larva < pupa < adult$, with a minimal sensitivity of at least 0.9 for all the classes. However, when the data was projected using $(pupa, adult)$, a slightly different order was obtained, with minimal class sensitivity of 0.94. The reported suborders of length three were acquired with a minimum sensitivity of 1 for each class. Similarly, the zebrafish transitions, from embryo to adult, were predicted to follow the expected sequence, with sensitivity 1 for all classes, when the class projection (3d, 1–2yrs) was used. Whereas employing different class projections, the predicted orders exhibit some discrepancies, treating nearby stages as substitutes; for instance, we frequently observe 3d overlapping with 5d, and 3mo overlapping with 1–2yrs.

The proposed approach was further applied on the two datasets pertaining pancreatic cancer, the human PDAC and mouse PanNET. The results displayed in Figure 7 were obtained with minimal class-wise sensitivity of 1. Here, we employed projecting pairs that describe remote stages of the process under consideration. The pairs (normal, PDAC), (PanIN-1, PDAC), (normal, PanIN-3) and (PanIN-1, PanIN-3) were examined for PDAC. For PanNET we investigated (NM, MET), (HI, MET), (NM, MLP) and (HI, MLP). Each of these class combinations returned partial orders of length not greater than three for PDAC and not greater than four for PanNET.

It can be noticed that in both scenarios, no orders, comprising transitions from normal tissue to the metastatic disease, were predicted following a fully continuous or linear sequence. The observed sequences are characterized by gaps, where certain precursor lesions are noticeably absent.

4.3 Validation of detected structures

To validate the ordinal structures and substructures detected by our approach, we also applied the detection method introduced in (Bellmann et al., 2022). Since the method proposed by Bellmann et al. (2022) is limited to detecting total orders, we utilized it as follows. First, the complete datasets were analyzed. Subsequently, we evaluated all data subsets that contained only samples from the classes that constitute the longest substructures detected by our architecture. The evaluation led to the following outcomes.

As with our current approach, for the synthetic dataset, no total order was detected. The longest substructure consisting of six classes was confirmed, which is $l_0 < l_1 < l_2 < l_3 < l_4 < l_5$. For *D. rerio*, the total order was confirmed, as we detected with the projection class pair (3d, 1–2yrs). For PanNET, no total order was detected. All nine suborders of length 4 were confirmed, which are depicted in Figure 7, i.e., (NM < HI < MLP < MET), (NM < AI < MLP < MET), (NM < TI < MLP < MET), (HI < AI < MLP < MET), (HI < TI < MLP < MET), (NM < HI < TI < MLP), (NM < HI < TI < MET), (NM < HI

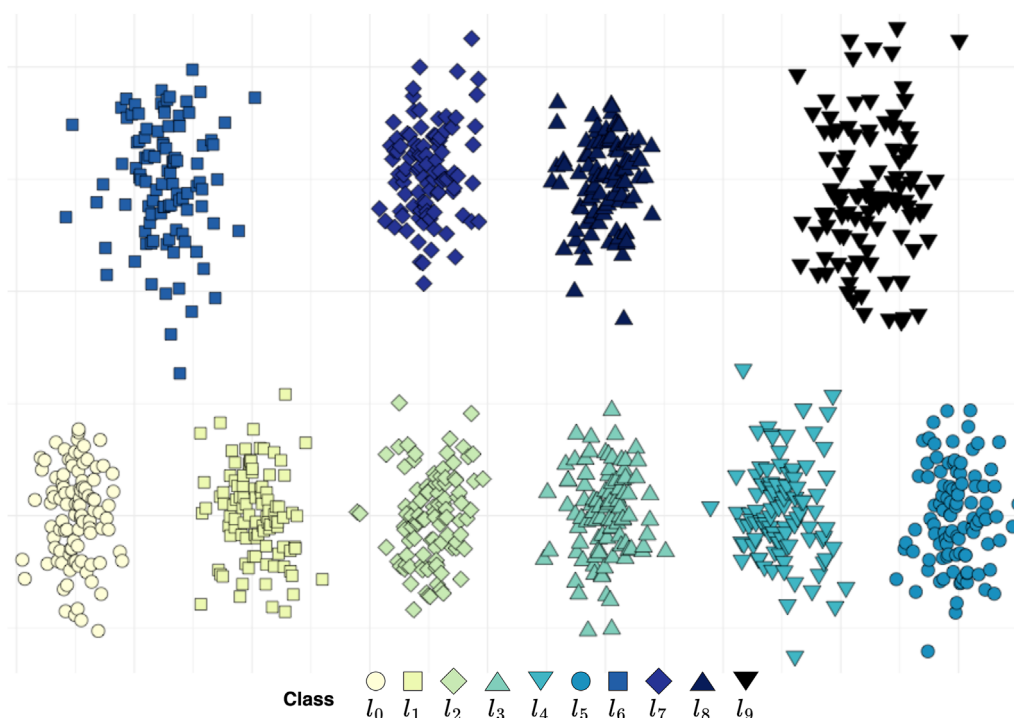


FIGURE 4

Synthetic two-dimensional data. Within the ten classes, no total order can be found, yet four suborders of length six are present: $(l_0 < \dots < l_5)$, $(l_0 < l_1 < l_2 < l_8 < l_4 < l_5)$, $(l_0 < l_1 < l_7 < l_3 < l_4 < l_5)$ and $(l_0 < l_1 < l_7 < l_8 < l_4 < l_5)$. In addition, 8 subsequences of length five can be likewise identified, resembling the previous ones where sequences $(l_0 < l_1)$ and $(l_4 < l_5)$ are replaced by classes l_6 and l_9 , respectively.

$< \text{AI} < \text{MLP}$), and $(\text{NM} < \text{HI} < \text{AI} < \text{MET})$. For the PDAC dataset, again no total order was found and the following suborders were confirmed (cf. Figure 7): $(\text{normal} < \text{PanIN-1} < \text{PDAC})$, $(\text{PanIN-1} < \text{PanIN-2} < \text{PDAC})$, and $(\text{PanIN-1} < \text{PanIN-3} < \text{PDAC})$. In contrast to our outcomes, for the data subset including the classes (normal, PanIN-2, PanIN-3), the approach of Bellmann et al. (2022) led to the unconventional order $\text{normal} < \text{PanIN-3} < \text{PanIN-2}$.

The most interesting case was observed for dataset *D. melanogaster*. While we were able to detect the conventional total structure $\text{embryo} < \text{larva} < \text{pupa} < \text{adult}$ with the projection pair (embryo, adult), no total order was detected with the approach proposed by Bellmann et al. (2022). To further analyze this phenomenon, we additionally evaluated all data subsets including three of the four total classes. The returned detected orders of length three were $(\text{larva} < \text{pupa} < \text{adult})$, $(\text{embryo} < \text{pupa} < \text{adult})$, $(\text{embryo} < \text{larva} < \text{adult})$, and $(\text{embryo} < \text{pupa} < \text{larva})$. The last suborder seems to lead to a failed detection of a total class structure.

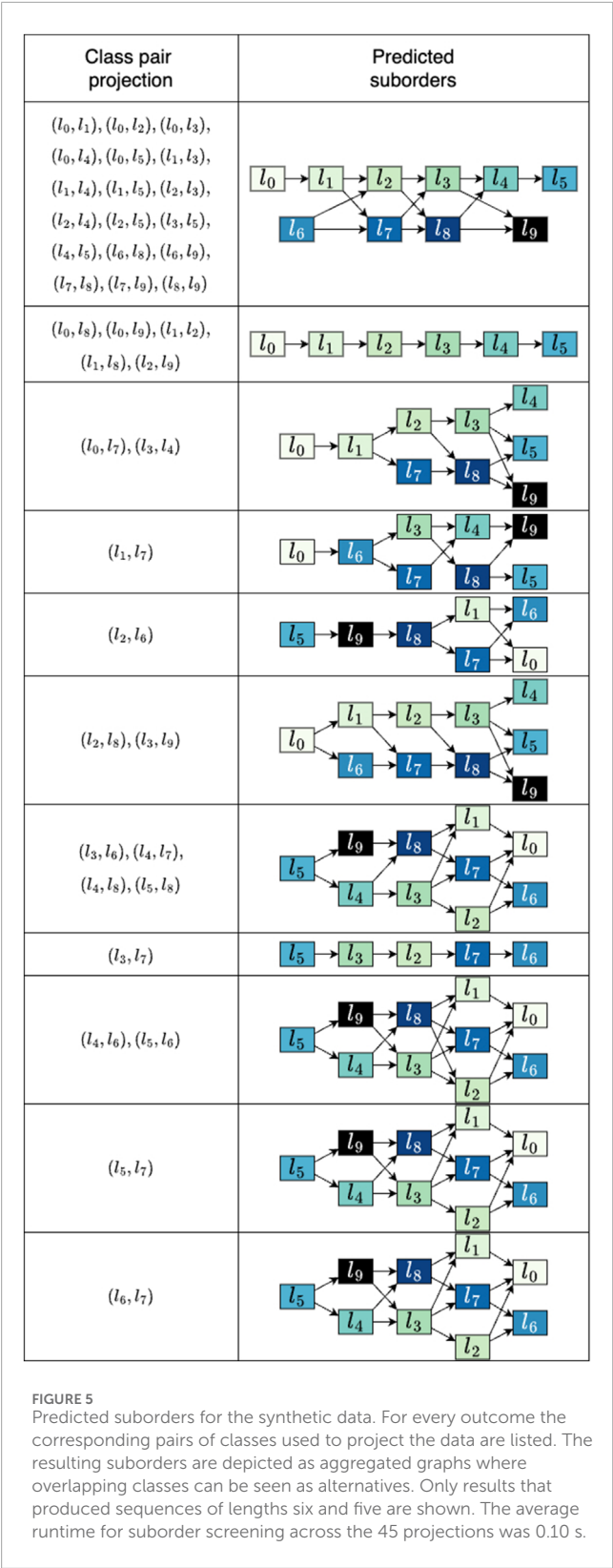
In summary, the comparison to the detection method introduced in (Bellmann et al., 2022) validated our detection architecture, emphasizing the benefit that our proposed approach is not limited to the analysis of total orders.

5 Discussion and conclusion

Uncovering ordinal correlations concealed within high-throughput data might significantly enhance our understanding of genetic alterations underlying various biological processes and assist

in predicting plausible disease progression. This paper introduces a methodology to retrieve univariate representations from high-throughput datasets and further analyze them using an advanced ordinal classification framework. This approach is especially suitable when examining intricate biological mechanisms concerning, for instance, cancer progressions such as pancreatic ductal adenocarcinomas (PDACs) or pancreatic neuroendocrine tumors (PanNETs). Alongside their unpredictable non-linear progressions, these tumors often exhibit heterogeneous staging among different patients, as well as within an individual (Jones et al., 2008; Raphael et al., 2017; Witkiewicz et al., 2015; Adamo et al., 2017), indicating the likely presence of distinct alternative developmental trajectories. In order to investigate these possibilities we integrate our novel directed threshold classifiers with the existent ordinal classifier cascades. The combination of these two techniques enables the detection of underlying ordinal substructures, which can be further aggregated to partial orders to reveal potential coexisting transition routes.

The approach used for projecting the data into a single-dimensional space plays a critical role in determining the effectiveness of the identified ordinal patterns. Specifically, we observed that selecting biologically distant class pairs for the initial binary separation results in a more pronounced separability of the categories throughout the entire progression. Although the approach could benefit from additional domain expertise concerning the definition of remote stages, it still proves effective, also in its absence. One option to choose an *effective* projection class pair, without focusing on the biological meaning of the classes, is



to conduct an exhaustive search over all possible class pairs and to select the two *most distant* classes, based on the provided feature space. If the biological order or possible (parallel) suborders are reflected in the provided feature space, the exhaustive search is

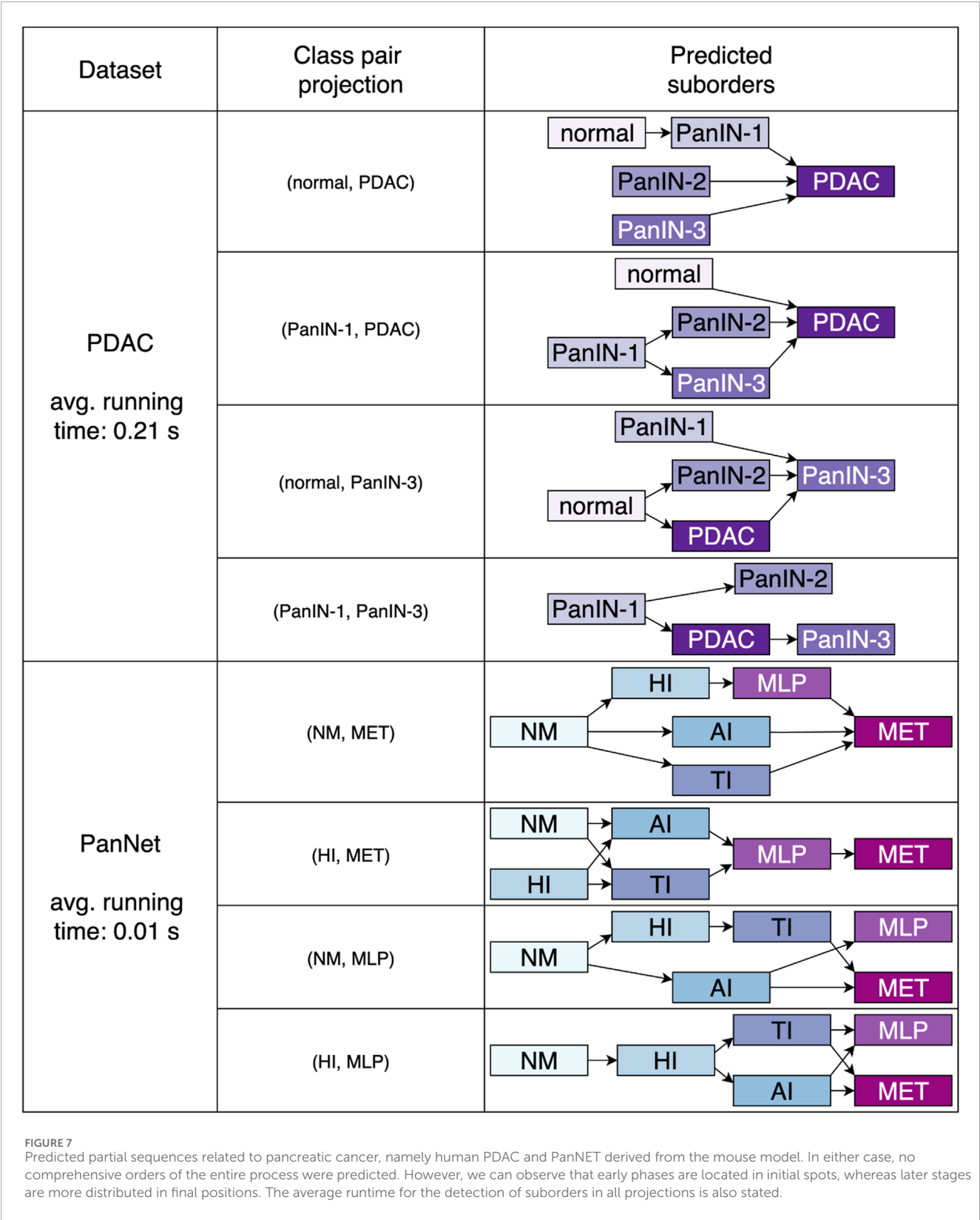
expected to lead to a meaningful initial projection class pair. Note that, despite choosing a well-founded data transformation to one dimension, ordinal detection may be influenced by class imbalance.

The method was validated on both, synthetic and biological datasets. When applied to the artificially generated dataset, engineered to include multiple suborders, the method accurately recovered all alternative sequences. Furthermore, we successfully rebuilt known linear stage orders in developmental data from *Drosophila melanogaster* and *Danio rerio*. These preliminary findings support the suggestion that employing projection pairs describing biologically distant stages in a specific developmental process, may more effectively direct the classifier in recognizing also intermediate phases, unlike using closely related stages that might hide certain transitions. Moreover, the results also prove that the methodology is suitable for detecting overall orders encompassing all classes, as well as suborders within data that lack an underlying total order.

Predicting the staging and progression becomes more challenging when investigating oncological datasets, such as PDACs and PanNETs (Buchholz et al., 2005; Ro et al., 2013; Chan et al., 2018; Mpilla et al., 2020). In these cases, the classifiers failed to recognize a uniform and stepwise course of the diseases from the onset to the ending phase. For the human pancreatic cancer, the pancreatic intraepithelial neoplasia of degree 1 (PanIN-1) as well as dysplasias of degrees 2 (PanIN-2) and 3 (PanIN-3) appear to be followed by PDAC. This observation is consistent with the current literature characterizing pancreatic carcinomas as mostly heterogeneous tumors with a complex evolution, whereby different tumor regions can develop independently of each other (Felsenstein et al., 2018). On the cellular level, PanINs arise from neoplastic transformation of normal cells like ductal, acinar, central acinar and normal stem cells in the exocrine part of the pancreas. Various molecular changes, as well as mutations in different signaling pathways (Hedgehog, Wnt, EGF, Notch and IL-17), contribute to varying degrees to the evolution of PanIN lesions in PDAC with a key role for Notch signalling (Pian et al., 2025). This leads to the formation of many subclonal populations, supporting the hypothesis that some malignancies might not follow a single linear progression model, but rather develop through multiple, parallel evolutionary routes (Notta et al., 2017; Wu et al., 2019). Previous transcriptional profiles analyses revealed a substantial difference between lesions and malignant pancreatic tumors, with the earliest lesions resembling more closely normal tissues (Buchholz et al., 2005). This is also evident in the arrangements that result from our detection technique. The research conducted by Notta et al. (2016) revealed that approximately 65% of PDAC tumors exhibit complex chromosomal rearrangements, including chromothripsis, a phenomenon in which chromosomes massively split and rejoin in a single event (Stephens et al., 2011). Multiple tumor suppressor genes, including TP53, CDKN2A, and SMAD4, can be simultaneously inactivated by this process, leading to rapid development and spread of tumors. These findings further challenge the conventional model of incremental genetic alterations in PDAC progression, suggesting that in some cases, the disease may advance rapidly due to such genomic failures. These molecular insights also align with the sudden onset of an advanced disease and the transition of duct lesions to invasive carcinoma that have been documented in clinical settings of certain patients (Hruban et al., 1999; Al-Sukhni et al.,

Dataset	Class pair projection	Predicted (sub)orders
D. melanogaster avg. running time: 0.006 s	(embryo, adult)	embryo → larva → pupa → adult
	(pupa, adult)	embryo → pupa → larva → adult
	(embryo, larva)	embryo → pupa → larva
	(embryo, pupa)	embryo → larva → pupa adult
	(larva, pupa)	larva → embryo → pupa adult
	(larva, adult)	embryo → pupa → adult larva
D. rerio avg. running time: 0.01 s	(3d, 1 – 2yrs)	3d → 5d → 10d → 3mo → 1-2yrs
	(3d, 3mo)	3d → 5d → 10d → 3mo 1-2yr
	(10d, 1 – 2yrs)	10d → 3d → 3mo → 1-2yr 5d
	(3d, 10d)	3d → 5d → 3mo → 10d 1-2yr
	(5d, 3mo), (5d, 1 – 2yrs)	3d → 10d → 3mo 5d → 1-2yr
	10d, 3mo,	10d → 3d → 3mo 5d → 1-2yr
	(5d, 10d)	3d → 3mo → 10d 5d → 1-2yr

FIGURE 6
Predicted overall and partial sequences for the two developmental datasets. The projecting class pairs that returned orders either matching the length of the expected order or one element shorter (assumed length – 1) are provided for both *Drosophila melanogaster* and *Danio rerio*. The average computation time for suborders detection in all projections is also provided.



2012). The observation of patients undergoing yearly magnetic resonance imaging screenings revealed that although imaging could detect small pancreatic tumors and cystic lesions, some participants still developed higher stage PDAC with minimal or no prior symptoms. Pancreatic neuroendocrine tumors can be clinically differentiated into functionally active and inactive types, and further subdivided into well-differentiated and poorly differentiated subgroups. Further subtyping of this clinically heterogeneous tumor

entity can be achieved by integrating molecular information that may be relevant to tumor development and progression (Shen et al., 2022). Despite the fact that PDAC and PanNETs are distinct entities, several studies highlight the role of chromatin remodeling and genomic alterations in pancreatic tumorigenesis, showing both similarities and differences between the two (De Wilde et al., 2012; Iacobuzio-Donahue et al., 2012; Jiao et al., 2011).

A valuable foundation for comprehending tumor heterogeneity was provided by examining the RIP1-TAG2 mouse model as a representation of human PanNETs (Sadanandam et al., 2015). The integration of transcriptomic and metabolic profiling across human and mouse models led to the identification of multiple tumor subtypes, each characterized by unique molecular and clinical features. This work reveals concurrent routes of PanNET carcinogenesis, exhibiting distinctive cells of origin that result in tumor islets and metastasis-like primary subtypes, strengthening the concept of non-linear development of these tumor types.

In conclusion, the approach we introduce offers a foundation for examining variability in the development of diseases, effectively unveiling underlying potential ordinal patterns. Additional research into intricate biological and pathological mechanisms, particularly understanding the distinct developmental routes in both PanNETs and PDAC may have significant implications for prognostic evaluations and tailored treatment plans. While the presented outcomes were obtained from relatively small datasets, further research will focus on external validation with larger sample cohorts, together with the analysis of additional technical modifications. An example could be to complement the OCC sensitivity by alternative OC measures, such as the weighted κ (Cohen, 1968) or Kendall's τ (Kendall, 1938).

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

AS: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing. PE-B: Formal Analysis, Methodology, Writing – original draft,

Writing – review and editing. AK: Writing – review and editing. HK: Conceptualization, Formal Analysis, Funding acquisition, Investigation, Project administration, Resources, Supervision, Writing – original draft, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. HK acknowledges funding from the German Science Foundation (DFG, SFB 1506, Aging at Interfaces, no. 450627322 and GRK 3012, KEMAI, no. 520750254) and the German Federal Ministry of Education and Research (BMFTR, Medical Informatics Initiative, project Private AIM, no. 01ZZ2316N).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Adamo, P., Cowley, C. M., Neal, C. P., Mistry, V., Page, K., Dennison, A. R., et al. (2017). Profiling tumour heterogeneity through circulating tumour dna in patients with pancreatic cancer. *Oncotarget* 8, 87221–87233. doi:10.18632/oncotarget.20250
- Al-Sukhni, W., Borgida, A., Rothenmund, H., Holter, S., Semotiuk, K., Grant, R., et al. (2012). Screening for pancreatic cancer in a high-risk cohort: an eight-year experience. *J. Gastrointest. Surg.* 16, 771–783. doi:10.1007/s11605-011-1781-6
- Arbeitman, M. N., Furlong, E. E., Imam, F., Johnson, E., Null, B. H., Baker, B. S., et al. (2002). Gene expression during the life cycle of drosophila melanogaster. *Science* 297, 2270–2275. doi:10.1126/science.1072152
- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2012). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi:10.1093/nar/gks1193
- Beadsmoore, C., and Screaton, N. (2003). Classification, staging and prognosis of lung cancer. *Eur. J. Radiology* 45, 8–17. doi:10.1016/s0720-048x(02)00287-5
- Bellmann, P., and Schwenker, F. (2020). Ordinal classification: working definition and detection of ordinal structures. *IEEE Access* 8, 164380–164391. doi:10.1109/access.2020.3021596
- Bellmann, P., Lausser, L., Kestler, H. A., and Schwenker, F. (2022). A theoretical approach to ordinal classification: feature space-based definition and classifier-independent detection of ordinal class structures. *Appl. Sci.* 12, 1815. doi:10.3390/app12041815
- Breivik, H., Borchgrevink, P.-C., Allen, S.-M., Rosseland, L.-A., Romundstad, L., Breivik Hals, E., et al. (2008). Assessment of pain. *Br. J. Anaesth.* 101, 17–24. doi:10.1093/bja/aen103

- Brody, J. (2009). Parallel routes of human carcinoma development: implications of the age-specific incidence data. *Nat. Preced.* 4, e7053. doi:10.1371/journal.pone.007053
- Buchholz, M., Kestler, H. A., Bauer, A., Böck, W., Rau, B., Leder, G., et al. (2005). Specialized dna arrays for the differentiation of pancreatic tumors. *Clin. Cancer Res.* 11, 8048–8054. doi:10.1158/1078-0432.ccr-05-1274
- Chan, C. S., Laddha, S. V., Lewis, P. W., Koletsy, M. S., Robzyk, K., Da Silva, E., et al. (2018). Atrx, daxx or men1 mutant pancreatic neuroendocrine tumors are a distinct alpha-cell signature subgroup. *Nat. Commun.* 9, 4158. doi:10.1038/s41467-018-06498-2
- Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull.* 70, 213–220. doi:10.1037/h0026256
- Cortés, J., Calvo, E., Vivancos, A., Perez-García, J., Recio, J. A., and Seoane, J. (2014). New approach to cancer therapy based on a molecularly defined cancer classification. *CA a Cancer J. Clin.* 64, 70–74. doi:10.3322/caac.21211
- Davis, M., O'Connell, T., Johnson, S., Cline, S., Merikle, E., Martenyi, F., et al. (2018). Estimating alzheimer's disease progression rates from normal cognition through mild cognitive impairment and stages of dementia. *Curr. Alzheimer Res.* 15, 777–788. doi:10.2174/1567205015666180119092427
- De Wilde, R. F., Edil, B. H., Hruban, R. H., and Maitra, A. (2012). Well-differentiated pancreatic neuroendocrine tumors: from genetics to therapy. *Nat. Rev. Gastroenterology and Hepatology* 9, 199–208. doi:10.1038/nrgastro.2012.9
- Felsenstein, M., Hruban, R. H., and Wood, L. D. (2018). New developments in the molecular mechanisms of pancreatic tumorigenesis. *Adv. Anatomic pathology* 25, 131–142. doi:10.1097/pap.0000000000000172
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7, 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x
- Förner, A., Díaz-González, Á., Liccioni, A., and Vilana, R. (2014). Prognosis prediction and staging. *Best Pract. and Res. Clin. Gastroenterology* 28, 855–865. doi:10.1016/j.bpg.2014.08.002
- Frank, E., and Hall, M. (2001). “A simple approach to ordinal classification,” in *Machine learning: ECML 2001: 12th european conference on machine learning Freiburg, Germany, September 5–7, 2001 proceedings 12* (Springer), 145–156.
- Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* 366, 883–892. doi:10.1056/nejmoa1113205
- Goh, W. W. B., Wang, W., and Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.* 35, 498–507. doi:10.1016/j.tibtech.2017.02.012
- Hadjistavropoulos, T., and Craig, K. D. (2002). A theoretical framework for understanding self-report and observational measures of pain: a communications model. *Behav. Res. Ther.* 40, 551–570. doi:10.1016/s0005-7967(01)00072-9
- Haefeli, M., and Elfering, A. (2006). Pain assessment. *Eur. Spine J.* 15, S17–S24. doi:10.1007/s00586-005-1044-x
- Hruban, R. H., Wilentz, R., Goggins, M., Offerhaus, G., Yeo, C., and Kern, S. (1999). Pathology of incipient pancreatic cancer. *Ann. Oncol.* 10, S9–S11. doi:10.1093/annonc/10.suppl_4.s9
- Iacobuzio-Donahue, C. A., Velculescu, V. E., Wolfgang, C. L., and Hruban, R. H. (2012). Genetic basis of pancreas cancer development and progression: insights from whole-exome and whole-genome sequencing. *Clin. Cancer Res.* 18, 4257–4265. doi:10.1158/1078-0432.ccr-12-0315
- Jiao, Y., Shi, C., Edil, B. H., De Wilde, R. F., Klimstra, D. S., Maitra, A., et al. (2011). Daxx/atrx, men1, and mtor pathway genes are frequently altered in pancreatic neuroendocrine tumors. *Science* 331, 1199–1203. doi:10.1126/science.1200609
- Jones, S., Zhang, X., Parsons, D. W., Lin, J. C.-H., Leary, R. J., Angenendt, P., et al. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 321, 1801–1806. doi:10.1126/science.1164368
- Kambhatla, N., and Leen, T. K. (1997). Dimension reduction by local principal component analysis. *Neural Comput.* 9, 1493–1516. doi:10.1162/neco.1997.9.7.1493
- Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika* 30, 81–93. doi:10.1093/biomet/30.1-2.81
- Lattke, R., Lausser, L., Müssel, C., and Kestler, H. A. (2015). “Detecting ordinal class structures,” in *Multiple classifier systems: 12th international workshop, MCS 2015, günzburg, Germany, June 29–July 1, 2015, proceedings 12* (Springer), 100–111.
- Lausser, L., Schäfer, L. M., Schirra, L.-R., Szekeley, R., Schmid, F., and Kestler, H. A. (2019). Assessing phenotype order in molecular data. *Sci. Rep.* 9, 11746. doi:10.1038/s41598-019-48150-z
- Lausser, L., Schäfer, L. M., Kühlwein, S. D., Kestler, A. M. R., and Kestler, H. A. (2020). Detecting ordinal subcascades. *Neural Process. Lett.* 52, 2583–2605. doi:10.1007/s11063-020-10362-0
- Lee, J.-S., Chu, I.-S., Heo, J., Calvisi, D. F., Sun, Z., Roskams, T., et al. (2004). Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. *Hepatology* 40, 667–676. doi:10.1002/hep.20375
- McInnes, L., and Healy, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. *Corr. abs/1802.03426*. doi:10.48550/arXiv.1802.03426
- Mpilla, G. B., Philip, P. A., El-Rayes, B., and Azmi, A. S. (2020). Pancreatic neuroendocrine tumors: therapeutic challenges and research limitations. *World J. Gastroenterology* 26, 4036–4054. doi:10.3748/wjg.v26.i28.4036
- Notta, F., Chan-Seng-Yue, M., Lemire, M., Li, Y., Wilson, G. W., Connor, A. A., et al. (2016). A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature* 538, 378–382. doi:10.1038/nature19823
- Notta, F., Hahn, S. A., and Real, F. X. (2017). A genetic roadmap of pancreatic cancer: still evolving. *Gut* 66, 2170–2178. doi:10.1136/gutjnl-2016-313317
- Olschwang, S., Hamelin, R., Laurent-Puig, P., Thuille, B., De Rycke, Y., Li, Y.-J., et al. (1997). Alternative genetic pathways in colorectal carcinogenesis. *Proc. Natl. Acad. Sci.* 94, 12122–12127. doi:10.1073/pnas.94.22.12122
- Pian, L.-L., Song, M.-h., Wang, T.-f., Qi, L., Peng, T.-L., and Xie, K.-p. (2025). Identification and analysis of pancreatic intraepithelial neoplasia: opportunities and challenges. *Front. Endocrinol.* 15, 1401829–1402024. doi:10.3389/fendo.2024.1401829
- Prigogine, I., and Nicolis, G. (1971). Biological order, structure and instabilities. *Q. Rev. Biophysics* 4, 107–148. doi:10.1017/s0033583500000615
- Raphael, B. J., Hruban, R. H., Aguirre, A. J., Moffitt, R. A., Yeh, J. J., Stewart, C., et al. (2017). Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* 32, 185–203.e13. doi:10.1016/j.ccell.2017.07.007
- Ro, C., Chai, W., Victoria, E. Y., and Yu, R. (2013). Pancreatic neuroendocrine tumors: biology, diagnosis, and treatment. *Chin. J. Cancer* 32, 312–324. doi:10.5732/cjc.012.10295
- Sadanandam, A., Wullschlegler, S., Lyssiotis, C. A., Gröttinger, C., Barbi, S., Bersani, S., et al. (2015). A cross-species analysis in pancreatic neuroendocrine tumors reveals molecular subtypes with distinctive clinical, metastatic, developmental, and metabolic characteristics. *Cancer Discov.* 5, 1296–1313. doi:10.1158/2159-8290.cd-15-0068
- Scharre, D. W. (2019). Preclinical, prodromal, and dementia stages of alzheimer's disease. *Pract. Neurol.* 15, 36–47.
- Seoane, J., and De Mattos-Arruda, L. (2014). The challenge of intratumour heterogeneity in precision medicine. *J. Intern. Med.* 276, 41–51. doi:10.1111/joim.12240
- Shen, X., Wang, X., Lu, X., Zhao, Y., and Guan, W. (2022). Molecular biology of pancreatic neuroendocrine tumors: from mechanism to translation. *Front. Oncol.* 12, 967071–2022. doi:10.3389/fonc.2022.967071
- Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., et al. (2011). Toward defining the preclinical stages of alzheimer's disease: recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's and Dementia* 7, 280–292. doi:10.1016/j.jalz.2011.03.003
- Stephens, P. J., Greenman, C. D., Fu, B., Yang, F., Bignell, G. R., Mudie, L. J., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 144, 27–40. doi:10.1016/j.cell.2010.11.055
- Tahami Monfared, A. A., Byrnes, M. J., White, L. A., and Zhang, Q. (2022). Alzheimer's disease: epidemiology and clinical progression. *Neurology Ther.* 11, 553–569. doi:10.1007/s40120-022-00338-8
- Toyama, R., Chen, X., Hawar, N., Aamar, E., Epstein, J., Reany, N., et al. (2009). Transcriptome analysis of the zebrafish pineal gland. *Dev. Dyn. Official Publ. Am. Assoc. Anatomists* 238, 1813–1826. doi:10.1002/dvdy.21988
- Traverso, G., Shuber, A., Levin, B., Johnson, C., Olsson, L., Schoetz Jr, D. J., et al. (2002). Detection of apc mutations in fecal dna from patients with colorectal tumors. *N. Engl. J. Med.* 346, 311–320. doi:10.1056/nejmoa012294
- Tu, Y., Stolovitzky, G., and Klein, U. (2002). Quantitative noise analysis for gene expression microarray experiments. *Proc. Natl. Acad. Sci.* 99, 14031–14036. doi:10.1073/pnas.222164199
- Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.* 9.
- Vapnik, V. N. (2000). *The nature of statistical learning theory, second edition. Statistics for engineering and information science.* Springer.
- Witkiewicz, A. K., McMillan, E. A., Balaji, U., Baek, G., Lin, W.-C., Mansour, J., et al. (2015). Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat. Commun.* 6, 6744. doi:10.1038/ncomms7744
- Wu, R.-C., Wang, P., Lin, S.-F., Zhang, M., Song, Q., Chu, T., et al. (2019). Genomic landscape and evolutionary trajectories of ovarian cancer precursor lesions. *J. Pathology* 248, 41–50. doi:10.1002/path.5219
- Yang, D., Jones, M. G., Naranjo, S., Rideout, W. M., Min, K. H. J., Ho, R., et al. (2022). Lineage tracing reveals the phylogenetics, plasticity, and paths of tumor evolution. *Cell* 185, 1905–1923.e25. doi:10.1016/j.cell.2022.04.015