Expert Systems With Applications 266 (2025) 126078

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

HLOB–Information persistence and structure in limit order books

Antonio Briola [a,*], Silvia Bartolucci [a], Tomaso Aste [a,b]

[a] Department of Computer Science, University College London, London, United Kingdom
[b] Systemic Risk Centre, London School of Economics, London, United Kingdom

ARTICLE INFO

Keywords:
Market microstructure
Limit order book
Econophysics
High frequency trading
Deep learning

ABSTRACT

We introduce a novel large-scale deep learning model for Limit Order Book mid-price changes forecasting, and we name it 'HLOB'. This architecture (i) exploits the information encoded by an Information Filtering Network, namely the Triangulated Maximally Filtered Graph, to unveil deeper and non-trivial dependency structures among volume levels; and (ii) guarantees deterministic design choices to handle the complexity of the underlying system by drawing inspiration from the groundbreaking class of Homological Convolutional Neural Networks. We test our model against 9 state-of-the-art deep learning alternatives on 3 real-world Limit Order Book datasets, each including 15 stocks traded on the NASDAQ exchange, and we systematically characterize the scenarios where HLOB outperforms state-of-the-art architectures. Our approach sheds new light on the spatial distribution of information in Limit Order Books and on its degradation over increasing prediction horizons, narrowing the gap between microstructural modeling and deep learning-based forecasting in high-frequency financial markets.

1. Introduction

Financial markets are complex environments. Their complexity stems from two main factors: (i) the interaction of a large number of agents pursuing heterogeneous goals at different time scales through the implementation of trading strategies designed to leverage asymmetric information; (ii) the emergence of self-organizing collective behaviors that do not result from the existence of any central controller and are therefore difficult to anticipate. The concurrence of these aspects contributes to the sporadic and limited-in-time persistence of inefficiencies that make the trading practice profitable. The analysis of existing inefficiencies and the forecasting of new ones is made possible by the mathematical and statistical modeling of the time series reflecting the financial market's behavior. The granularity of these time series widely varies depending on the goal of the analysis, and, in the high-frequency case (i.e., the scenario we are mainly interested in), it can be order-driven with a resolution up to the nanosecond (LOBSTER Data, 2023).

Indeed, the majority of modern financial exchanges store order-level updates in data structures known as Limit Order Books (LOBs). At each point in time, in a given automated exchange, these data structures contain a snapshot of the standing intentions of market participants to buy or sell different amounts (or volumes) of an asset at a given price. Such trading intentions, which are defined in jargon as 'orders', can be of different types (i.e., market orders, limit orders, and cancellation orders) and their flux (i.e., incoming or outgoing) is generally managed by computerized systems exploiting a *FIFO* (first-in, first-out) mechanism to establish execution's priority (Bouchaud, Bonart, Donier, & Gould, 2018; Briola, Bartolucci, & Aste, 2024; Briola, Turiel, & Aste, 2020; Briola, Turiel, Marcaccioli, Cauderan, & Aste, 2021). The timing of accessing information contained in LOBs guarantees asymmetric levels of information to market participants. At the finest-grained information's exploitation level, we refer to High-Frequency Trading (HFT) to indicate the strategies that gain an edge through speed, allowing certain traders to act on information not yet accessible to others (Lehalle & Laruelle, 2018). HFT strategies exploit market's microstructure imperfections to the detriment of other traders, triggering a predator–prey dynamic with other actors (Farmer & Skouras, 2013). HFT has been prominent in the financial landscape since 2005 (Isichenko, 2021). Despite being object of criticism and regulatory scrutiny since its introduction, it has been demonstrated that this practice's reliance on various levels of market data, rather than external information, contributes to noise generation, thereby preserving unpredictability in stock price movements (Bouchaud, Farmer, & Lillo, 2009).

The difficulty in handling the inherent complexity expressed by HFT systems and the availability of large amounts of data, has fostered the development of deep learning models as a solution to the related modeling and forecasting tasks. Over recent years, increasingly sophisticated solutions have emerged, with some of them evolving towards

* Correspondence to: Department of Computer Science, University College London, 66-72 Gower Street, WC1E 6EA, London, United Kingdom.
E-mail addresses: antonio.briola.20@ucl.ac.uk (A. Briola), s.bartolucci@ucl.ac.uk (S. Bartolucci), t.aste@ucl.ac.uk (T. Aste).

https://doi.org/10.1016/j.eswa.2024.126078
Received 8 July 2024; Received in revised form 16 November 2024; Accepted 3 December 2024
Available online 12 December 2024

the creation of informed architectures that meticulously incorporate LOBs' components in the definition of derived features. Although many scientific investigations proved the potential of these approaches, there is still a noticeable disconnection between theoretical results and their real-world practicability (Prata et al., 2023). Furthermore, the recent research work by Briola et al. (2024) highlights that the effectiveness of these methods considerably varies depending on the stocks' unique microstructural characteristics. More specifically, the authors show that microstructural properties of stocks exposed to higher trading risks (i.e., the so-called 'small-tick stocks') impose sparser LOB structures, highly undermining the ability of deep learning architectures to model their hidden dynamics effectively. On the contrary, the microstructural properties of stocks exposed to lower trading risks (i.e., the so-called 'large-tick stocks') impose more compact LOB structures, facilitating deep learning architectures in effectively processing the underlying information.

The contribution of our paper is threefold:

1. We introduce 'HLOB', a novel, large-scale deep learning architecture that exploits the class of the Homological Convolutional Neural Networks (Briola, Wang, Bartolucci, & Aste, 2023; Wang, Briola, & Aste, 2023) to superimpose a dependency structure among LOB volume levels and model deeper and non-trivial relationships among them[1].
2. We show that the exploitability of the informational content encoded in the spatial structure imposed by our deep learning architecture is limited in time and the velocity of its degradation is highly dependent on the stocks' microstructural properties.
3. We test our model against 9 state-of-the-art deep learning alternatives on 3 real-world LOB datasets, each including 15 stocks traded on the NASDAQ exchange. Our findings highlight the difficulty in finding a model that consistently outperforms the others; hence, we provide the guidelines for selecting a model based on factors such as the desired level of interpretability, the specific forecasting horizon, and the available infrastructure.

The rest of the paper is organized as follows. In Section 2, we provide an overview of the essential scientific works describing (i) the functioning of LOB dynamics; (ii) the main architectures proposed in the past to solve LOB-related forecasting tasks; and (ii) the intuition behind Information Filtering Networks and the class of Homological Convolutional Neural Networks. In Section 3, we present an overview of the datasets used in our experiments. In Section 4, we provide technical insights into the HLOB model and the framework used for its training and validation. In Section 5, we present the results of our experiments, while in Section 6, we wrap up our findings, providing a comprehensive description of the power and weaknesses of our model compared to the existing ones, with an overview on open challenges in the field.

2. Related work

In this Section, we provide the essential references to (i) understand the operational mechanics of the LOBs; (ii) become familiar with existing models designed to identify microstructural alphas; and (iii) grasp the theoretical foundations of the HLOB model. It is essential to notice that our investigation spans three distinct research fields: (i) market microstructure; (ii) deep learning; and (iii) network science. We do not claim to cover the entire related literature, but, for each domain, we selectively reference the works that are critically relevant to our research, equipping the reader with the foundational tools required to master the content of this paper.

[1] The code to reproduce all the experiments is available at https://github.com/FinancialComputingUCL/LOBFrame/tree/main.

2.1. Limit order book

Most modern financial exchanges utilize electronic systems to record and match the trading intentions of market participants. These systems are centered on a data structure called 'Limit Order Book' (LOB), which is unique for each security traded on a given exchange and provides immediate access to real-time supply and demand in the visible market. Participants on the same side of the market (whether buying or selling) compete with each other while concurrently opposing those on the opposite side; the buyers want to buy cheaper, and the sellers want to sell at a higher price, but the two sides ultimately need each other to make trades happen. The LOB is, hence, subject to updates (or ticks) that occur at irregular time intervals. These events reflect changes in the market and are constrained by predefined adjustments: (i) the tick size (θ) for price adjustments; and (ii) the lot size (ψ) for volume changes.[2] Updates are made possible through the submission of new orders. Based on their direction, they can be bid (buy) or ask (sell) orders; based on their aggressive or passive attitude, they can be market or limit orders. A market order expresses the necessity to buy or sell a certain amount of a given asset at the current best available price on the opposite side of the LOB; it is typically subject to higher transaction fees. A limit order expresses an intention to buy or sell a quantity of an asset at a price that is more advantageous to the one quoted on the best level of the LOB[3]; it populates a queue in one of the deeper levels of the LOB, it does not have any guarantee to be executed and is typically subject to lower transaction fees. Cancellations represent a third class of orders; they delete active limit orders and are typically not subject to transaction fees.

Temporally, the LOB is structured as stacked snapshots reflecting the tick-by-tick evolution of the market, and takes the form of a multivariate time-series $\mathbb{L} \in \mathbb{R}^{T\times 4L}$, where T is the history length, and L is the number of levels.[4] Spatially, a LOB record can be represented as:

$$\mathbb{L}(\tau) = \{p_{\ell}^{\text{ask}}(\tau), v_{\ell}^{\text{ask}}(\tau), p_{\ell}^{\text{bid}}(\tau), v_{\ell}^{\text{bid}}(\tau)\}_{\ell=1}^{L}, \tag{1}$$

where $p_{\ell}^{\text{ask/bid}}(\tau)$ is the ask/bid price at level $\ell \in L$ and $v_{\ell}^{\text{ask/bid}}(\tau)$ is the volume on the same level $\ell \in L$. The mid-price m_{τ} of a stock at time τ is defined as the average between the best ask price (i.e., $p_1^{\text{ask}}(\tau)$) and the best bid price (i.e. $p_1^{\text{bid}}(\tau)$), $m_{\tau} = \frac{p_1^{\text{ask}}(\tau)+p_1^{\text{bid}}(\tau)}{2}$. The bid–ask spread σ_{τ} of the stock at time τ is defined as the difference between the best ask price and the best bid price, $\sigma_{\tau} = p_1^{\text{ask}}(\tau) - p_1^{\text{bid}}(\tau)$.

The level-based representation in Eq. (1) is convenient from the perspective of human understanding of the functioning of a LOB. However, it suffers a significant drawback from an automated learning standpoint: indeed, there is no guarantee of homogeneous spatial separation between consecutive price levels. It is worth noticing that, when exacerbated by specific stock's microstructural properties (see the research works by Bouchaud et al. (2018), Briola et al. (2024), Sirignano and Cont (2021)), such heterogeneity in the spatial distribution of LOB data sensibly reduces the ability of specific classes of deep-learning models (e.g., Convolutional Neural Networks) in the micro alphas' discovering process (Wu, Mahfouz, Magazzeni, & Veloso, 2021).

[2] The value of θ and ψ depend on the exchange. In the NASDAQ exchange, which is the source of the data used in the current research work (see Section 3), $\theta = \$0.01$ and $\psi = 1$.

[3] A LOB is organized into price/volume levels. On the bid side, standing intentions to buy different quantities of a financial security are organized in a descending order (i.e., the first level contains the orders to be executed at the highest price among the quoted ones); on the ask side, standing intentions to sell different quantities of a financial security are organized in an ascending order (i.e., the first level contains the orders to be executed at the lowest price among the quoted ones).

[4] The dimensionality here $4L$ because, for each level, we register the corresponding ask price, ask volume, bid price, and bid volume.

2.2. Deep learning for limit order book forecasting

The difficulty in handling the complexity expressed by LOBs and the related data abundance has fostered the development of deep learning algorithms to solve related modeling and forecasting tasks. Among them, we are particularly interested in architectures designed to forecast the direction of mid-price changes at a high-frequency resolution. Foundational contributions in the field are offered by Passalis et al. (2017), Sirignano (2019), Sirignano and Cont (2021), Tsantekidis et al. (2017a, 2017b). These studies introduce the use of Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997), Convolutional Neural Network (CNN) (LeCun, Bengio, & Hinton, 2015), and Bag-of-Features (BoF) (O'Hara & Draper, 2011) architectures as viable approaches for the forecasting task. Subsequently, these modules served as core components in more complex architectures. This trend emerged in the works of Zhang, Yao, Sun, and Tay (2019) and Tsantekidis et al. (2020), where the authors utilize convolutional filters to capture the spatial structure of the LOB, as well as LSTM modules to capture long-term time dependencies, and in the works by Passalis, Tefas, Kanniainen, Gabbouj, and Iosifidis (2020) and Tran, Passalis, Tefas, Gabbouj, and Iosifidis (2022), where the authors enrich the BoF paradigm for LOB forecasting through the introduction of the attention mechanism (Vaswani et al., 2017). Other relevant works are the ones by Tran, Iosifidis, Kanniainen, and Gabbouj (2018), Tran, Kanniainen, Gabbouj, and Iosifidis (2021) and Shabani, Tran, Kanniainen, and Iosifidis (2023), Shabani, Tran, Magris, Kanniainen, and Iosifidis (2022), where the authors propose an architecture that incorporates the idea of bi-linear projection as well as of attention to focus on crucial temporal and spatial information embedded in LOBs.

Concerning the integration of attention mechanisms in LOB forecasting attempts, it is worth mentioning the work by Guo and Chen (2023), where the authors introduce a dual-stage temporal attention mechanism to repeatedly highlight the most valuable time-dimension information, and the works by Kisiel and Gorse (2022), Wallbridge (2020), Zhang, Lim, and Zohren (2021), which use transformer-based architectures to accomplish similar forecasting tasks. Lastly, it is relevant to mention the research by Briola et al. (2024, 2020), Kolm, Turiel, and Westray (2023), Kolm and Westray (2024), Lucchese, Pakkanen, and Veraart (2022), where the authors critically assess the efficacy of methodologies mentioned previously in this section to understand their effectiveness under different evaluation conditions.

2.3. Information filtering networks & homological (convolutional) neural networks

One of the main contributions of this paper is the introduction of HLOB, a novel large-scale deep learning model for mid-price change forecasting at a high-frequency resolution. This architecture is designed to capture and exploit complex dependencies at deeper LOB levels. This approach overcomes traditional CNN-LSTM models (e.g., DeepLOB (Zhang, Zohren, & Roberts, 2019)), which only capture dependencies between consecutive LOB levels, being inadequate to fully handle the inherent complexity of the underlying system.

The key theoretical prior behind HLOB is represented by Information Filtering Networks (IFNs) (Aste, Di Matteo, & Hyde, 2005; Barfuss, Massara, Di Matteo, & Aste, 2016; Mantegna, 1999; Massara, Di Matteo, & Aste, 2016; Tumminello, Aste, Di Matteo, & Mantegna, 2005). IFNs are an effective tool to represent and model dependency structures among variables characterizing complex systems through the instruments of network science, while imposing strict topological constraints (e.g., being a tree or a planar graph) and optimizing global properties (e.g., the model's likelihood) (Aste, 2022). The filtering process can be performed in many different ways; historically, the three main examples of IFNs have been (i) the Minimum Spanning Tree (MST) (West et al., 2001); (ii) the Planar Maximally Filtered Graph (PMFG) (Aste & Di Matteo, 2006; Tumminello, Di Matteo, Aste, & Mantegna, 2007); and (iii) the Triangulated Maximally Filtered Graph (TMFG) (Massara, Di Matteo, & Aste, 2017). In this paper, we are mainly interested in the latter. The TMFG captures higher-order relationships among up to four variables per clique being planar[5] and chordal[6], and maximizes the likelihood of the underlying system by deterministically joining in a recursive way covariates expressing the highest similarity (Briola & Aste, 2022; Massara et al., 2017). This class of IFNs also inspired the groundbreaking class of Deep Neural Networks at the core of HLOB: the Homological Convolutional Neural Networks (HCNNs) (Briola et al., 2023). This architecture, which has an archetype in the simpler class of Homological Neural Network (Wang et al., 2023), is entirely data-centric and leverages the power of convolutions to take advantage of the topological priors in the TMFG (Briola et al., 2023). An in-depth description of the building process of a TMFG, of an HCNN, and of the HLOB originating from them, is provided in Sections 4.1 and 4.2.

3. Data

We analyze 15 stocks from 6 sectors and 13 industries, all listed on the NASDAQ exchange. The chosen dataset was originally proposed by Briola et al. (2024) and contains only assets maintaining a large- (i.e., 10B-200B) to -mega (i.e., $\geq$ 200B) capitalization on a 3-year analysis period spanning from January 2017 to December 2019. Stock-related information are summarized in Table 1, where the assets are organized into 3 groups based on their tick size.

The *first group* (i.e., CHTR, GOOG, GS, IBM, MCD, NVDA) contains 'small-tick stocks' (i.e., the stocks characterized by $\langle\sigma\rangle \geq 3\theta$, where $\langle\sigma\rangle$ indicates the average bid–ask spread). The *second group* (i.e., AAPL, ABBV, PM) contains 'medium-tick stocks' (i.e., the stocks characterized by $1.5\theta \lesssim \langle\sigma\rangle \lesssim 3\theta$). The *third group* (i.e., BAC, CSCO, KO, ORCL, PFE, VZ) contains large-tick stocks (i.e., the stocks characterized by $\langle\sigma\rangle \lesssim 1.5\theta$). An in-depth description of the effectiveness of this classification in capturing stocks- and class-related microstructural effects is available in the original research work (Briola et al., 2024).

For each stock, high-resolution, tick-by-tick LOB data obtained from the LOBSTER provider (LOBSTER Data, 2023) are employed. For each trading day, we use a LOB characterized by $L = 10$ price and volume levels for both the bid and ask sides (see Eq. (1)). As outlined in Table 2, for each year, we allocate 40 days for training, 5 days for validation, and 10 consecutive days for testing. Notably, the training days are chosen to form a sequence where most of the entries are consecutive, with only few exceptions. Indeed, the 5 days of validation are randomly selected from the same period characterizing the training set. This choice guarantees greater robustness in the validation step, and it is made possible by the 5-days feature-wise rolling window z-score standardization procedure, which prevents any data leakage (Briola et al., 2024). The raw LOB data are processed in accordance with the rigorous pipeline initially proposed by Lucchese et al. (2022) and subsequently refined by Briola et al. (2024).

Consistently with the work by Briola et al. (2024), we study the predictability of mid-price changes' direction[7] at 3 different horizons

[5] A graph is said to be planar if it can be embedded in a sphere without edges crossing.

[6] A graph is said to be chordal if all cycles made of four or more vertices have a chord which reduces the cycle to a set of triangles. A chord is defined as an edge that is not part of the cycle but connects two vertices of the cycle itself.

[7] We decide to use the simple difference in mid-prices to gain higher control over the amplitude of the change at different time horizons, preserving, at the same time, the stationarity property of the resulting time series. Many alternatives have been proposed as target variables in the literature (e.g., Lucchese et al. (2022), Ntakaris, Magris, Kanniainen, Gabbouj, and Iosifidis (2018), Tsantekidis et al. (2017a), Zhang, Zohren, and Roberts (2019)). All of them are based on the usage of the log-return as fundamental quantity, and apply different smoothing methods to mitigate the strong fit between labels and actual

Table 1
Overview of the stocks used in the paper. For each asset, we report the ticker, the extended name, the sector, the industry and the capitalization during 2017, 2018 and 2019. To determine stocks' sector and industry affiliation, we follow the taxonomy proposed by the NASDAQ exchange (NASDAQ, 2023). To determine the stock's capitalization, we rely on the data provided by companiesmarketcap.com (companiesmarketcap.com, 2024).

Stock symbol	Stock name	Sector	Industry	Capitalization (2017)	Capitalization (2018)	Capitalization (2019)
CHTR	Charter Communications, Inc.	Telecommunications	Cable & Other Pay Television Services	$83.94 B	$64.21 B	$101.85 B
GOOG	Alphabet, Inc.	Technology	Computer Software: Programming, Data Processing	$729.45 B	$723.55 B	$921.13 B
GS	Goldman Sachs Group, Inc.	Finance	Investment Bankers/Brokers/Service	$96.09 B	$61.43 B	$79.86 B
IBM	International Business Machines Corporation	Technology	Computer Manufacturing	$142.03 B	$101.44 B	$118.90 B
MCD	McDonald's Corporation	Consumer Discretionary	Restaurants	$137.21 B	$136.21 B	$147.47 B
NVDA	NVIDIA Corporation	Technology	Semiconductors	$117.26 B	$81.43 B	$144.00 B
AAPL	Apple, Inc.	Technology	Computer Manufacturing	$860.88 B	$746.07 B	$1.287 T
ABBV	AbbVie, Inc.	Health Care	Biotechnology: Pharmaceutical Preparations	$154.39 B	$136.33 B	$130.94 B
PM	Philip Morris International, Inc.	Health Care	Medicinal Chemicals and Botanical Products	$164.09 B	$103.78 B	$132.39 B
BAC	Bank of America Corporation	Finance	Major Banks	$307.91 B	$238.25 B	$311.20 B
CSCO	Cisco Systems, Inc.	Telecommunications	Computer Communications Equipment	$189.34 B	$194.81 B	$203.45 B
KO	Coca-Cola Company	Consumer Staples	Beverages (Production/Distribution)	$195.47 B	$202.08 B	$236.89 B
ORCL	Oracle Corporation	Technology	Computer Software: Prepackaged Software	$195.72 B	$162.03 B	$169.94 B
PFE	Pfizer, Inc.	Health Care	Biotechnology: Pharmaceutical Preparations	$215.89 B	$249.54 B	$216.82 B
VZ	Verizon Communications, Inc.	Telecommunications	Telecommunications Equipment	$215.92 B	$232.30 B	$253.93 B

Table 2
Basic structure of the datasets used during the training, validation and test stage. For each year, for the training and test set, we report the starting and the ending day (both included in the analysis), while, for the validation set, we report all the dates explicitly. It is worth noting that weekends and public holidays are not trading days and, consequently, do not belong to any of the datasets.

Year	Training		Validation	Test	
	from	to	days	from	to
2017	03–13	05–22	03–23, 04–05, 04–13, 04–18, 05–02	05–23	06–06
2018	08–09	10–18	08–15, 08–16, 09–19, 09–26 10–03	10–19	11–01
2019	06–04	08–13	06–14, 06–27, 07–08, 07–10, 07–24	08–14	08-27

(i.e., $H\Delta_\tau \in \{10, 50, 100\}$) when such a movement is larger than or equal to θ. The labeling step is consequently defined as follows:

$$\begin{cases} (m_{\tau+\Delta\tau} - m_\tau) \leq -\theta \;\rightarrow\; -1 \;\rightarrow\; \text{Down}\,, \\ -\theta < (m_{\tau+\Delta\tau} - m_\tau) < +\theta \;\rightarrow\; 0 \;\rightarrow\; \text{Stable}\,, \\ (m_{\tau+\Delta\tau} - m_\tau) \geq +\theta \;\rightarrow\; 1 \;\rightarrow\; \text{Up}\,, \end{cases} \tag{2}$$

where $\theta = \$0.01$ is the tick size on the NASDAQ exchange and m_τ is the mid-price at tick time τ. It is worth noticing that horizons are always defined in terms of LOB updates (which are unevenly spaced), while physical time is never used.

Fig. 1 reports the normalized average class distribution across the training, validation, and test set over the 3-year analysis period of investigation, for $H\Delta_\tau \in \{10, 50, 100\}$[8]. Generally speaking, it is always possible to detect imbalances; their evolution across horizons is, however, different for different groups of stocks. For *small-tick stocks*, it is possible to observe a rough balance at $H\Delta_\tau \in \{10\}$. In contrast, an increasingly pronounced imbalance towards the extreme classes (i.e., -1 and 1) is observed moving to longer prediction horizons, $H\Delta_\tau \in \{50, 100\}$. In a less evident way, the same trend can also be observed for *medium-tick stocks*. This mitigation depends on the nature of the class of stocks, which contains both assets behaving more similarly to small-tick stocks and assets behaving more similarly to large-tick stocks. For this latest class of stocks (i.e., *large-tick stocks*), we observe a pattern diametrically opposite to the one described for small-tick stocks. Indeed, at $H\Delta_\tau \in \{10\}$ we observe a strong imbalance towards the central class, which is mitigated by moving to longer prediction horizons until reaching a rough balance at horizon $H\Delta_\tau \in \{100\}$. All the imbalances described earlier in the paragraph are handled in the training stage through the usage of balanced data-loaders. However, their effects will remain partially visible in the test set's data.

4. Methods

The HLOB model is centered on two primary mechanisms: (i) exploiting the informational content of topological priors in an IFN as input for a tailored version of Homological Convolutional Neural Networks (HCNNs) (Briola et al., 2023), which in turn handles the dependency structures among LOB's spatial components (i.e., volume levels); and (ii) employing an LSTM module to capture long-term temporal patterns. A complete description of this system requires (i) a preliminary discussion on the process to distillate the necessary information to build the TMFG; and (ii) a detailed description of the required modifications to the original HCNN architecture to make it suitable for processing LOB inputs.

4.1. The TMFG's building process

The building block of the HCNN and, consequently, of the HLOB architecture, is represented by an arbitrary IFN encoding higher-order dependency structures among variables in the underlying system. In this study, in line with the work by Briola et al. (2023), we choose the Triangulated Maximally Filtered Graph (TMFG) (Massara et al., 2017). As a *first step*, we process LOB data by removing price levels from the bid and ask side, focusing solely on volume-related data[9]. Formally, we reduce the dimension of each LOB snapshot from $\mathbb{L}(\tau) \in \mathbb{R}^{4L}$ to $\mathbb{L}(\tau) \in \mathbb{R}^{2L}$. This choice is needed to ensure homogeneity in the information used to build the IFN. Volume levels are inherently discrete, with the minimum tradable quantity set by the exchange's ψ parameter (see Section 2.1), and minor variations from consecutive LOB updates can introduce a non-negligible level of noise. To mitigate this effect, we categorize volumes into equally spaced bins. The number of bins is optimized on the training and validation set, and remains constant across stocks characterized by different microstructural properties (i.e., small-, medium-, and large-tick stocks). The size of the bins is calculated for each stock individually, and across all volume levels for each training day.

As a *second step*, for each stock and for each day in the training set, we calculate the pairwise mutual information (MI) between volume levels, obtaining positive and symmetric $(2L \times 2L)$ similarity matrices

prices. While these methods are academically acceptable, their practicability is questionable as they are more tailored towards tracking mid-price trends than immediate fluctuations, thereby offering limited control on tick-by-tick changes crucial for developing high-frequency trading strategies.

[8] For a more detailed investigation of class imbalances across the training, validation, and test set, the reader is referenced to work by Briola et al. (2024).

[9] At this stage, we consider unscaled data-only.

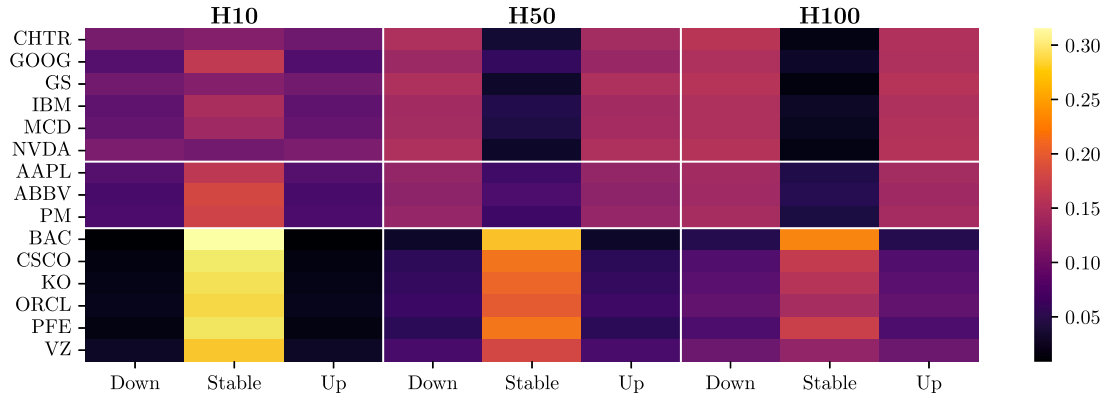

Fig. 1. Normalized average class distribution across the training, validation, and test set over a 3-year analysis period at $H\Delta_\tau \in \{10, 50, 100\}$. Class imbalances vary significantly across horizons and stock groups. For small-tick stocks, near-balance is observed at $H\Delta_\tau \in \{10\}$, with increasingly pronounced imbalance at longer prediction horizons. Large-tick stocks show strong class imbalance towards the central class at $H\Delta_\tau \in \{10\}$, transitioning to near-balance at $H\Delta_\tau \in \{100\}$. Medium-tick stocks exhibit less pronounced trends, with mixed behaviors.

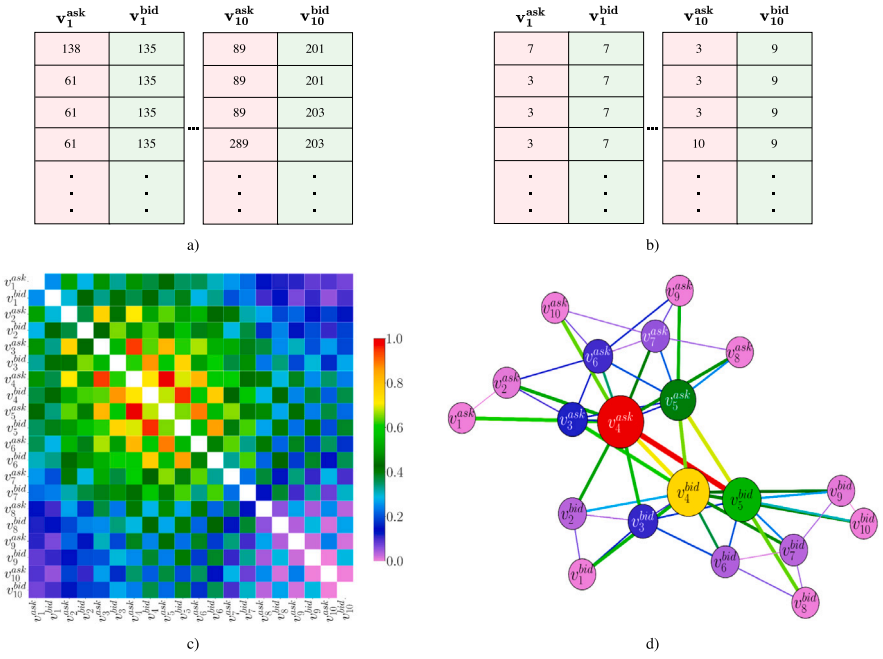

Fig. 2. Schematic representation of the TMFG's building process: (a) we start from a simplified version of the LOB containing only volumes data; (b) we mitigate the noise affecting the LOB by categorizing volumes into bins of uniform size; (c) we compute the pairwise MI between volume levels; and (d) we build the TMFG using the MI matrix as input. We remark that in the proposed graph representation, both nodes' and edges' color/dimension depend on their betweenness centrality. The color bar remains consistent for both the MI matrix and the corresponding TMFG representation. It is worth noticing that the TMFG captures not only local interactions (i.e., interactions between consecutive volume levels on the same side of the LOB, or across volumes on the same level but opposite sides of the LOB), but also deeper and non-trivial dependency structures among volume levels on the same and opposite sides of the LOB.

(i.e., MI matrices). It is worth noticing that the reliability of the MI computation is strengthened through a bootstrapping process applied on a daily basis on LOB data. For each stock, the final MI matrix is derived by averaging daily MI matrices in the training set.

As a *third step*, stock-related TMFGs are computed by using average MI matrices as similarity matrices. We remark that, given the multivariate system $\mathbb{L}$, our primary goal is to estimate the multivariate probability density function $\widetilde{f}(\mathbb{L}|\mathcal{G}^*)$ with representation structure $\mathcal{G}^*$ that best describes the true and unknown $f(\mathbb{L})$. From an information theoretic perspective, the learning of an optimal network representation $\mathcal{G}^*$ consists of minimizing the Kullback–Leibler divergence (D_{KL}) (Kullback & Leibler, 1951) between $f(\mathbb{L})$ and $\widetilde{f}(\mathbb{L}|\mathcal{G})$, and, consequently, the cross-entropy (H) of the underlying system:

$$\begin{aligned}\mathcal{G}^* &\Rightarrow \arg\min_{\mathcal{G}} D_{KL}(f(\mathbb{L}) \parallel \widetilde{f}(\mathbb{L}|\mathcal{G})) \\ &\Rightarrow \arg\min_{\mathcal{G}} \cancel{\mathbb{E}_f(\log f(\mathbb{L}))} - \mathbb{E}_f(\log \widetilde{f}(\mathbb{L}|\mathcal{G})) \\ &\Rightarrow \arg\min_{\mathcal{G}}(H(\mathbb{L}|\mathcal{G}))\end{aligned} \tag{3}$$

The term $\mathbb{E}_f(\log f(\mathbb{L}))$ in Eq. (3) is independent from $\mathcal{G}$ and therefore its value is irrelevant to the purpose of discovering the optimal representation network. The second term, $-\mathbb{E}(\log \widetilde{f}(\mathbb{L}|\mathcal{G}))$ (notice the minus), instead, depends on $\mathcal{G}$ and must be minimized. It is the estimate of the entropy of the multivariate system under analysis and corresponds

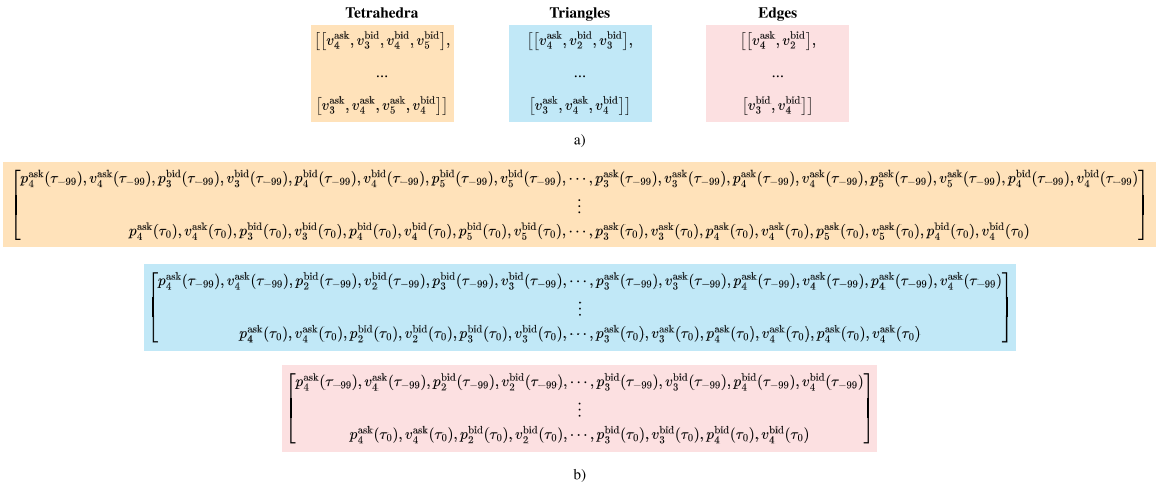

Fig. 3. This diagram illustrates the sequence of steps transitioning (a) from the output of the TMFG building process (b) to the input of the HLOB model. To construct the TMFG, we exclusively utilize volume levels from the LOB, forming a network characterized by three topological structures: tetrahedra, triangles, and edges. To prepare the inputs for the HLOB model, we perform two main tasks: (i) for each timestamp in the input's temporal dimension, we flatten each of the aforementioned sets; (ii) we incorporate the corresponding price levels' data into each representative of these three new input sets. Note that there is a direct mapping between the colors used in this Figure and the ones used later to highlight the inputs of the HLOB model in Fig. 4.

to the so-called cross-entropy (H). This minimization problem can be incrementally solved by joining the system's disconnected parts sharing the largest MI, which is exactly what the TMFG algorithm does (see Massara et al. (2017)).

It is worth noticing that the double auction mechanism underlying the LOB allows orders' placement dynamics that may manifest their effects over heterogeneous time scales (Bouchaud et al., 2018, 2009; Jain, Firoozye, Kochems, & Treleaven, 2024a, 2024b; Jain, Muzy, Kochems, & Bacry, 2024; Lehalle & Laruelle, 2018). Our approach captures local interactions and deeper, non-trivial dependencies that unfold across LOB's levels and sides at different temporal horizons (see Fig. 2). The TMFG building process guarantees this advantage and, as we explain in Section 4.2, allows the HCNN to overcome a major limitation characterizing existing models, which instead exploit non-linear transformations to describe local interactions across consecutive price and volume levels incrementally.

4.2. From HCNN to HLOB

From each TMFG computed as described in Section 4.1, we isolate the realizations of 3 simplicial families: (i) maximal cliques with size 4 (i.e., 3-dimensional simplices or *tetrahedra*); (ii) maximal cliques with size 3 (i.e., 2-dimensional simplices or *triangles*); (iii) maximal cliques with size 2 (i.e., 1-dimensional simplices or *edges*).

These three higher-order structures are sufficient to capture all the dependencies described by the chosen IFN. Given that the number of observed volume levels is constant across different stocks and trading days, we can deterministically compute (i) the shape of the vector of tetrahedra (17×4); (ii) the shape of the vector of triangles (52×3); and (iii) the shape of the vector of edges, (54×2). All of them serve as input for the HLOB model[10], which, however, is designed to handle not only the *spatial* dynamics captured by the TMFG, but also the *temporal* dynamics of the LOB. In this sense, as model's input, consistently with the work by Zhang, Zohren, and Roberts (2019), we also use a history window of 100 LOB's updates[11].

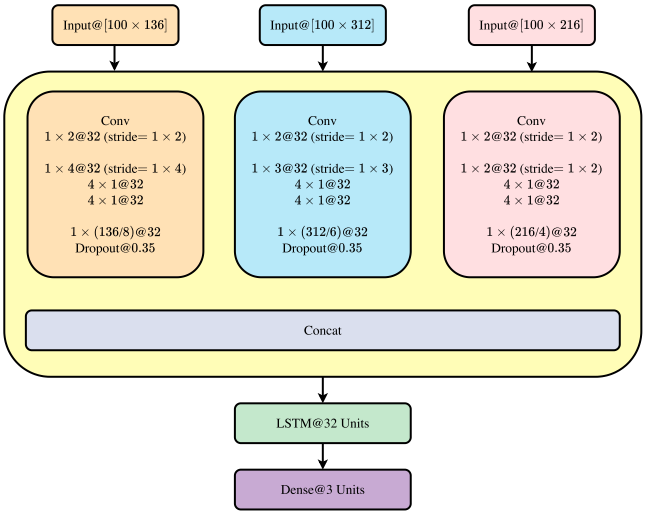

Fig. 4. Visual overview of the HLOB model's operational framework. Note that there is a direct mapping between the colors used to denote the inputs of the HLOB model here and the colors used to represent the three categories of topological priors derived from a TMFG in Fig. 3.

As described in Section 4.1, we use only the volume levels in the building process of the TMFG. However, price levels carry significant information that cannot be ignored. For this reason, we include them in the HLOB's building stage: for each timestamp constituting the input's historical dimension, we flat the vector of tetrahedra, triangles, and edges, and, for each volume level, we insert the corresponding price level. This transformation is schematically depicted in Fig. 3, where we show the flattening step for the 3 sets of simplicial families constituting the average TMFG, and the insertion of the price levels for each timestamp in the historical dimension of the HLOB's input. This operation produces 3 conceptually new 2D input vectors: (i) one of size (100×136) (i.e., the re-shaped vector of tetrahedra); (ii) one of size (100×312) (i.e., the re-shaped vector of triangles); and (iii) one of size (100×216) (i.e., the re-shaped vector of edges). Each vector is separately passed as input to one head of the HLOB model.

For each head, the size of the *first convolutional filter* is (1×2) with a stride of (1×2). As described in the work by Zhang, Zohren, and

[10] At this stage, we use scaled data-only (see Section 3).

[11] We remark the existence of strong designing analogies between the HLOB model and the DeepLOB one (Zhang, Zohren, & Roberts, 2019), which, indeed, represents an archetype for the architecture introduced in this research paper. For this reason, in Section 5, we will systematically discuss the comparison between the forecasting performances of these two models.

Roberts (2019), this first layer summarizes the information between the price and the volume $\{p^s_\ell, v^s_\ell\}^{s\in\{\text{ask, bid}\}}_{\ell\in L}$ at level ℓ and side s of the LOB. At the same time, the stride prevents parameter sharing between geographically (but not logically) consecutive inputs. The number of parameters corresponding to this operation equals 96 for each of the three heads of the architecture. The *second convolutional layer* captures the relationships between components of a single realization of each simplicial family: (i) in the case of tetrahedra, between nodes composing each 3−dimensional simplex (i.e., 4-cliques); (ii) in the case of triangles, between nodes composing each 2−dimensional simplex (i.e., 3-cliques); (iii) in the case of edges, between nodes composing each 1−dimensional simplex (i.e., 2-cliques). A stride equal to (1×4) for tetrahedra, (1×3) for triangles, and (1×2) for edges, one more time, prevents parameter sharing between components of the same simplicial family. Here, the number of parameters for the convolutional operation involving tetrahedra is equal to 12 384, the number of parameters for the convolutional operation involving triangles is equal to 11 360, while the number of parameters involving edges is equal to 10 336. The *third convolutional layer* captures the relationships between components of each simplicial family. The size of the convolutional filter is $(1\times\Omega)$, where Ω is the cardinality of each original set of simplexes: $136/8 = 17$ in the case of tetrahedra, $312/6 = 52$ in the case of triangles, and $216/4 = 54$ in the case of edges. This further level of convolution is proved to be effective (see the work by Briola et al. (2023)) in capturing information that is not necessarily related in the original network representation, but that can positively affect the characterization of the true but unknown $f(\mathbb{L})$ (see Eq. (3)). Since relationships modeled in this layer do not directly stem from the structure of the underlying TMFG, for each head of the HLOB, we apply a dropout with a rate of 0.35. The number of parameters for this convolutional layer equals 17 440 in the case of tetrahedra, 53 280 in the case of triangles, and 55 328 in the case of edges. After these three layers of convolution, the dimension of each head's feature map is (100×1). These outputs are concatenated and passed through an *LSTM module* to capture long-term temporal dependencies. The activation of an LSTM unit is fed back to itself, and the memory of past activations is kept with a separate set of weights, so the temporal dynamics of input features can be effectively modeled. The number of parameters of this additional layer is equal to 16 640. Lastly, the *output layer* consists of a linear layer with a number of outputs equal to the number of classes. The model returns the logits for increased numerical stability while associated probabilities are computed in a separate stage.

4.3. Experimental settings

We test the HLOB architecture against 6 state-of-the-art (SOTA) models in LOB mid-price changes forecasting: (i) CNN1 (Tsantekidis et al., 2017a); (ii) CNN2 (Tsantekidis et al., 2020); (iii) DLA (Guo & Chen, 2023); (iv) BinBTabl (Tran et al., 2021); (v) BinCTabl (Tran et al., 2021); (vi) DeepLOB (Zhang, Zohren, & Roberts, 2019). All these models were proposed in the scientific literature between 2017 and 2022, and later systematically organized in the review paper by Prata et al. (2023). We also test our model against 2 pure transformer-based architectures for time-series forecasting that we adapt for LOB mid-price changes forecasting: (i) Transformer (Vaswani et al., 2017); (ii) iTransformer (Liu et al., 2023). Finally, as an additional benchmark model, we combine the power of Transformers and CNNs in the LobTransformer architecture, which takes inspiration from the work of Wallbridge (2020), and is proposed here in a revised version. As pointed out in the work by Briola et al. (2024), the majority of these architectures suffer from a fundamental drawback in the science domain: the original code is not provided, severely compromising the results' reproducibility. For the first set of models described above, results discussed in the current paper are obtained by exploiting the code provided by Prata et al. (2023). All the other architectures are implemented from scratch. All models are included in the 'LOBFrame' (Briola et al., 2024) pipeline to simplify their execution, while guaranteeing the highest reproducibility standards. A summary of the benchmark models is reported in Table 3.

When possible, model-specific hyper-parameters are inherited from the work by Prata et al. (2023), while optimal weights are learned by minimizing the categorical cross-entropy loss using mini-batches of size 32 (Zhang, Zohren, & Roberts, 2019). The mini-batches sampling procedure differs for the training, validation, and test sets. During training, the (sub)-sampling is random and balanced. From each trading day (see Table 2), we detect the number of samples for the least represented class, and (i) if this value is $\geq$ 5000, then we sample 5000 random representatives for each of the three classes (see Eq. (2)), otherwise, (ii) if this value is $<$ 5000, we sample a number of random representatives for each class equal to the number of samples for the least represented class. During validation and test stages, we still sample mini-batches with a size of 32, but they are always sequential and cover the totality of data in the two sets. In line with the related literature (Zhang, Zohren, & Roberts, 2019), all models are trained for a maximum number of epochs equal to 100. Training halts if the validation loss fails to drop by at least 0.003 units over a span of 15 consecutive epochs. We use a modified version of the Adam optimizer (Kingma & Ba, 2014) with decoupled weight decay (Loshchilov & Hutter, 2017), commonly known as 'AdamW'. Following the latest applied research findings (Brown et al., 2020; Karpathy, 2024), we use a learning rate equal to 6×10^{-5}, a β_1 decay rate equal to 0.90, and a β_2 decay rate equal to 0.95. As described in the work by Briola et al. (2024), the choice of values for these parameters is determined by the training pipeline described above.

All the models considered in this paper are coded in Python using the PyTorch deep learning library (Paszke et al., 2019). Experiments are run on the University College London Computer Science Department's High-Performance Computing Cluster (UCL CS HPC Cluster, 2023). Given the 15 stocks in Table 1, and knowing that for each of the 3 years we challenge the 10 models described in Table 3 on 3 prediction horizons, we obtain that the number of year-wise experiments is equal to 450. Consequently, the total number of executed experiments is equal to 1 350 for a cumulative GPU runtime of 7 192 hours, 20 minutes, and 31 seconds. To accomplish the task, we used 10 different GPU models: (i) NVIDIA A100 80GB PCIe (21 experiments); (ii) NVIDIA A100-PCIE-40 GB (6 experiments); (iii) NVIDIA GeForce GTX 1080 Ti (362 experiments); (iv) NVIDIA GeForce RTX 2080 Ti (439 experiments); (v) NVIDIA GeForce RTX 4090 (187 experiments); (vi) NVIDIA RTX 6000 Ada Generation (139 experiments); (vii) NVIDIA TITAN X (Pascal) (96 experiments); (viii) NVIDIA TITAN Xp (60 experiments); (ix) Tesla V100-PCIE-16GB (10 experiments); (x) Tesla V100-PCIE-32GB (30 experiments).

5. Results

We present the results of our analysis (i) evaluating the effectiveness of the models introduced in Section 4.3 in predicting the direction of mid-price changes; and (ii) examining the HLOB behavior to unveil intricate patterns into the LOB levels' structural dependencies. In all the experiments, we assess the behavior for the three classes of stocks (i.e., small-, medium- and large-tick stocks) at $H\Delta_\tau \in \{10, 50, 100\}$. This approach enables us to examine the models' performances across different scenarios, thereby linking their effectiveness to the microstructural characteristics of the stocks.

5.1. Comparison of model performances

We investigate models' effectiveness in predicting mid-price change direction through 3 key metrics: (i) the F1 score; (ii) the Matthews Correlation Coefficient (MCC) (Gorodkin, 2004); and (iii) the probability of correctly executing a round-trip transaction (p_T)[12]. We report

Table 3
We report a summary of three main characteristics of benchmark models: (i) original code availability; (ii) model's number of trainable parameters; and (iii) model's inference time in milliseconds. The original code is not provided for 5 out of 6 of the models having a direct reference in the literature. BinBTabl is the most parsimonious among benchmark models with a number of trainable parameters equal to 6.6×10^3, while LobTransformer is the less parsimonious one with a number of trainable parameters equal to 2.0×10^6. The model with the lowest inference time is CNN1 (i.e., 0.07 ms, while the model with the highest inference time is LobTransformer (i.e., 0.29 ms).

	Tsantekidis et al. (2017a) CNN1 (2017)	Tsantekidis et al. (2020) CNN2 (2020)	Guo and Chen (2023) DLA (2022)	Vaswani et al. (2017) Transformer (2017)	Liu et al. (2023) iTransformer (2023)	Briola et al. LobTransformer (2024)	Tran et al. (2021) BinBTabl (2021)	Tran et al. (2021) BinCTabl (2021)	Zhang, Zohren, and Roberts (2019) DeepLOB (2019)	Briola et al. HLOB (2024)
original code availability	✗	✗	✗	–	–	–	✗	✗	✓	–
n. trainable parameters	3.5×10^4	2.8×10^5	2.2×10^5	1.1×10^5	1.1×10^5	2.0×10^6	6.6×10^3	2.2×10^4	1.4×10^5	1.8×10^5
inference time (ms)	0.07	0.14	0.15	0.16	0.15	0.29	0.19	0.13	0.16	0.16

Table 4
Models' performances at $H\Delta_\tau = 10$. For each deep learning architecture we report three key metrics: (i) the F1 score; (ii) the MCC; and (iii) the p_T. For each stock, we highlight the best performing model (green), the second-best performing model (blue) and the worst performing alternative (red); a model is considered superior to the others if the sum of the 3 performance metrics is maximal.

	H10																											
	cnn1			cnn2			dla			transformer			itransformer			lobtransformer			binbtabl			binctabl			deeplob			hlob		
	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T
CHTR	0.39	0.11	0.06	0.38	0.09	0.05	0.39	0.11	0.04	0.40	0.12	0.06	0.35	0.05	0.04	0.31	0.07	0.04	0.42	0.15	0.06	0.43	0.16	0.06	0.39	0.10	0.05	0.43	0.17	0.06
GOOG	0.42	0.16	0.04	0.42	0.16	0.04	0.39	0.13	0.03	0.41	0.15	0.04	0.27	0.04	0.04	0.44	0.18	0.05	0.45	0.18	0.08	0.46	0.20	0.08	0.45	0.19	0.04	0.46	0.21	0.05
GS	0.36	0.10	0.09	0.29	0.06	0.06	0.38	0.09	0.08	0.38	0.12	0.10	0.31	0.05	0.02	0.16	0.00	0.00	0.40	0.15	0.09	0.41	0.15	0.10	0.34	0.10	0.08	0.41	0.17	0.12
IBM	0.36	0.09	0.12	0.36	0.08	0.08	0.35	0.08	0.11	0.35	0.11	0.11	0.30	0.06	0.02	0.30	0.05	0.05	0.36	0.10	0.14	0.37	0.10	0.14	0.38	0.11	0.13	0.40	0.13	0.14
MCD	0.37	0.08	0.10	0.35	0.08	0.08	0.38	0.09	0.10	0.38	0.10	0.11	0.31	0.04	0.02	0.28	0.04	0.01	0.39	0.11	0.12	0.40	0.11	0.12	0.41	0.12	0.11	0.41	0.13	0.13
NVDA	0.31	0.06	0.07	0.24	0.00	0.00	0.33	0.07	0.06	0.36	0.08	0.10	0.24	0.02	0.03	0.22	0.02	0.00	0.41	0.13	0.13	0.41	0.13	0.14	0.34	0.08	0.09	0.40	0.12	0.14
AAPL	0.42	0.16	0.13	0.39	0.14	0.09	0.41	0.14	0.12	0.41	0.16	0.13	0.35	0.09	0.07	0.39	0.17	0.10	0.41	0.15	0.15	0.42	0.16	0.16	0.43	0.17	0.15	0.42	0.18	0.15
ABBV	0.38	0.13	0.12	0.39	0.12	0.10	0.39	0.13	0.11	0.40	0.15	0.13	0.31	0.06	0.03	0.33	0.10	0.04	0.36	0.13	0.13	0.37	0.14	0.13	0.39	0.13	0.13	0.42	0.18	0.14
PM	0.35	0.08	0.09	0.39	0.10	0.09	0.37	0.08	0.08	0.36	0.09	0.09	0.29	0.02	0.02	0.28	0.04	0.04	0.36	0.10	0.12	0.36	0.11	0.13	0.36	0.12	0.13	0.39	0.13	0.13
BAC	0.43	0.23	0.04	0.38	0.21	0.06	0.38	0.23	0.04	0.44	0.28	0.05	0.36	0.18	0.09	0.46	0.29	0.05	0.45	0.27	0.06	0.45	0.28	0.07	0.46	0.30	0.07	0.47	0.32	0.06
CSCO	0.47	0.29	0.08	0.50	0.29	0.09	0.47	0.28	0.07	0.48	0.29	0.08	0.41	0.19	0.11	0.47	0.27	0.08	0.45	0.28	0.08	0.44	0.27	0.07	0.49	0.30	0.08	0.50	0.33	0.08
KO	0.47	0.26	0.08	0.47	0.27	0.10	0.46	0.28	0.09	0.47	0.28	0.09	0.39	0.17	0.10	0.48	0.30	0.10	0.45	0.28	0.10	0.43	0.27	0.10	0.48	0.28	0.10	0.49	0.31	0.10
ORCL	0.47	0.27	0.10	0.45	0.26	0.09	0.45	0.26	0.10	0.48	0.30	0.11	0.38	0.16	0.07	0.48	0.31	0.11	0.46	0.27	0.10	0.44	0.26	0.10	0.49	0.32	0.11	0.48	0.32	0.11
PFE	0.43	0.24	0.09	0.42	0.24	0.09	0.43	0.25	0.09	0.44	0.25	0.09	0.36	0.17	0.11	0.45	0.27	0.09	0.47	0.29	0.09	0.46	0.28	0.10	0.46	0.27	0.09	0.49	0.32	0.10
VZ	0.47	0.23	0.08	0.42	0.17	0.07	0.47	0.26	0.09	0.47	0.27	0.10	0.39	0.16	0.10	0.46	0.25	0.10	0.45	0.26	0.10	0.41	0.24	0.10	0.49	0.28	0.11	0.46	0.28	0.10

the results of this analysis in Tables 4, 5, and 6, highlighting the best performing model (green), the second-best performing model (blue) and the worst performing alternative (red). For each stock, a model is considered superior to the others if the sum of the 3 performance metrics is maximal. Year-wise metrics are computed, and, for each horizon $H\Delta_\tau \in \{10, 50, 100\}$, only the average value is provided.

Looking at Table 4, we notice that, at $H\Delta_\tau = 10$, HLOB outperforms SOTA alternatives in the 73.3% of cases. For small-tick stocks, it is the best-performing model in 4/6 scenarios (i.e., CHTR, GS, IBM, MCD); in the case of GOOG, it is the second-best alternative, while in the case of NVDA, it is the third-best alternative. For medium-tick stocks, HLOB is the best-performing model in 3/3 scenarios (i.e., AAPL, ABBV, PM), while, for large-tick stocks, it is the best-performing option in 4/6 cases (i.e., BAC, CSCO, KO, PFE) and the second-best alternative in the remaining 2 scenarios (i.e., ORCL and VZ). The HLOB average F1 score is equal to 0.42 for small-tick stocks, 0.41 for medium-tick stocks, and 0.48 for large-tick stocks. The average MCC is equal to 0.16 for small- and medium-tick stocks, and to 0.33 for large-tick stocks. The average p_T is equal to 0.11 for small-tick stocks, 0.14 for medium-tick stocks, and 0.09 for large-tick stocks. Focusing on inter-models' dynamics, we observe that, for small- to medium-tick stocks, performances are very similar for all the 3 evaluation metrics except for iTransformer and LobTransformer (which are the worst-performing alternatives). For large-tick stocks, instead, we observe that also the worst-performing models, even showing a considerable distance from the best-performing alternative in traditional machine-learning metrics' realizations (i.e., F1 score and MCC), present competitive realizations in the case of p_T. Comparing HLOB performances with DeepLOB ones, we observe that (i) the average gain in F1 score is equal to 0.03 for small-tick stocks, 0.02 for medium-tick stocks, and 0.003 for large-tick stocks; (ii) the average gain in MCC is equal to 0.04 for small-tick stocks, 0.02 for medium-tick stocks, and 0.02 for large-tick stocks; (iii) the average gain in p_T is equal to 0.02 for large-tick stocks, 0.01 for medium-tick stocks, and 0.00 for large-tick stocks.

Looking at Table 5, we notice that, at $H\Delta_\tau = 50$, HLOB model outperforms SOTA alternatives in the 60% of cases (10% less than what happens at $H\Delta_\tau = 10$). For small-tick stocks, it is the best-performing model in 1/6 scenarios (i.e., IBM); in the case of GS and MCD, it is the second-best alternative, while in all the other cases (i.e., CHTR, GOOG, and NVDA), it is the third-best alternative. For medium-tick stocks, HLOB is the best-performing model in 3/3 scenarios (i.e., AAPL, ABBV, PM), while, for large-tick stocks, it is the best-performing model in 5/6 cases (i.e., BAC, CSCO, KO, ORCL, VZ), being the second-best alternative in the case of PFE. The HLOB average F1 score is equal to 0.36 for small-tick stocks (with a percentage decrease of 16.66% compared to the realization at $H\Delta_\tau = 10$), 0.40 for medium-tick stocks (with a percentage decrease of 2.50% compared to the realization at $H\Delta_\tau = 10$), and 0.58 for large-tick stocks (with a percentage increase of 17.24% compared to the realization at $H\Delta_\tau = 10$). The average MCC

[12] The p_T metric was firstly introduced by Briola et al. (2024) to describe the probability od correctly executing round-trip transactions. It is defined as $p_T = \frac{CT}{PT+TT-CT}$. PT is the *number of potential transactions* (a transaction happens when one is able to open a position and then close it); we use the term 'potential' because transactions are counted on the targets' set. TT is the *number of executed transactions*; it is computed in the same manner as PT, but on the predictions' set. CT is the *number of correctly executed transactions*; it counts how many times a transaction executed on the predictions' set has a correspondence in the targets' set. Being a probability measure, p_T takes values between 0 and 1.

Table 5
Models' performances at $H\Delta_\tau = 50$. For each deep learning architecture we report three key metrics: (i) the F1 score; (ii) the MCC; and (iii) the p_T. For each stock, we highlight the best performing model (green), the second-best performing model (blue) and the worst performing alternative (red); a model is considered superior to the others if the sum of the 3 performance metrics is maximal.

	H50																											
	cnn1			cnn2			dla			transformer			itransformer			lobtransformer			binbtabl			binctabl			deeplob			hlob		
	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T
CHTR	0.34	0.07	0.03	0.33	0.03	0.02	0.36	0.09	0.04	0.36	0.09	0.05	0.29	0.03	0.04	0.22	0.01	0.00	0.38	0.15	0.05	0.38	0.13	0.05	0.36	0.06	0.03	0.36	0.09	0.05
GOOG	0.31	0.06	0.03	0.35	0.06	0.03	0.30	0.06	0.05	0.37	0.12	0.05	0.29	0.03	0.04	0.31	0.07	0.03	0.42	0.17	0.07	0.43	0.17	0.07	0.37	0.09	0.05	0.42	0.16	0.06
GS	0.30	0.03	0.04	0.25	0.00	0.03	0.29	0.04	0.05	0.32	0.04	0.06	0.26	0.01	0.03	0.22	0.00	0.00	0.32	0.11	0.08	0.35	0.11	0.08	0.29	0.03	0.06	0.34	0.09	0.08
IBM	0.30	0.04	0.06	0.29	0.01	0.04	0.33	0.05	0.07	0.34	0.07	0.07	0.25	0.02	0.02	0.23	0.03	0.00	0.36	0.07	0.10	0.34	0.07	0.10	0.28	0.03	0.04	0.37	0.08	0.09
MCD	0.32	0.05	0.06	0.29	0.04	0.04	0.31	0.06	0.07	0.32	0.06	0.06	0.28	0.02	0.03	0.23	0.01	0.00	0.36	0.08	0.09	0.36	0.08	0.09	0.32	0.04	0.05	0.36	0.07	0.08
NVDA	0.26	0.02	0.03	0.17	0.00	0.00	0.34	0.04	0.06	0.28	0.05	0.05	0.17	0.01	0.01	0.23	0.01	0.00	0.37	0.12	0.09	0.37	0.12	0.09	0.20	0.01	0.01	0.30	0.06	0.07
AAPL	0.35	0.09	0.11	0.37	0.09	0.06	0.38	0.09	0.11	0.39	0.11	0.10	0.26	0.04	0.02	0.32	0.09	0.04	0.40	0.11	0.12	0.40	0.11	0.12	0.36	0.11	0.07	0.40	0.12	0.11
ABBV	0.35	0.09	0.08	0.34	0.08	0.05	0.36	0.10	0.11	0.36	0.10	0.09	0.24	0.03	0.01	0.25	0.04	0.01	0.37	0.11	0.11	0.38	0.11	0.11	0.32	0.08	0.06	0.40	0.12	0.10
PM	0.35	0.07	0.09	0.33	0.06	0.05	0.35	0.06	0.09	0.37	0.07	0.09	0.26	0.02	0.02	0.27	0.02	0.01	0.37	0.09	0.11	0.37	0.09	0.10	0.36	0.07	0.10	0.39	0.09	0.10
BAC	0.59	0.39	0.09	0.55	0.39	0.08	0.53	0.36	0.08	0.58	0.41	0.09	0.41	0.24	0.08	0.55	0.33	0.07	0.59	0.40	0.11	0.59	0.40	0.09	0.61	0.44	0.10	0.62	0.47	0.09
CSCO	0.55	0.34	0.14	0.55	0.34	0.12	0.54	0.31	0.15	0.57	0.36	0.13	0.41	0.18	0.11	0.53	0.31	0.10	0.52	0.33	0.15	0.53	0.33	0.13	0.58	0.37	0.13	0.60	0.40	0.16
KO	0.53	0.30	0.13	0.51	0.29	0.11	0.56	0.34	0.15	0.56	0.34	0.13	0.43	0.20	0.11	0.57	0.35	0.14	0.54	0.34	0.13	0.55	0.35	0.13	0.56	0.34	0.14	0.59	0.39	0.15
ORCL	0.53	0.30	0.15	0.51	0.30	0.14	0.52	0.30	0.15	0.54	0.33	0.15	0.40	0.17	0.12	0.52	0.30	0.14	0.52	0.29	0.15	0.51	0.29	0.15	0.56	0.34	0.15	0.57	0.36	0.16
PFE	0.53	0.32	0.13	0.53	0.32	0.10	0.54	0.34	0.12	0.53	0.33	0.13	0.42	0.20	0.10	0.55	0.35	0.11	0.55	0.34	0.13	0.56	0.35	0.13	0.57	0.38	0.11	0.56	0.37	0.12
VZ	0.50	0.27	0.13	0.46	0.24	0.13	0.47	0.26	0.13	0.51	0.30	0.15	0.39	0.14	0.11	0.53	0.30	0.14	0.51	0.28	0.16	0.51	0.28	0.15	0.50	0.28	0.12	0.52	0.31	0.15

is equal to 0.09 for small-tick stocks (with a percentage decrease of 77.78% compared to the realization at $H\Delta_\tau = 10$), 0.11 for medium-tick stocks (with a percentage decrease of 45.45% compared to the realization at $H\Delta_\tau = 10$), and to 0.38 for large-tick stocks (with a percentage increase of 13.15% compared to the realization at $H\Delta_\tau = 10$). The average p_T is equal to 0.07 for small-tick stocks (with a percentage decrease of 57.14% compared to the realization at $H\Delta_\tau = 10$), 0.10 for medium-tick stocks (with a percentage decrease of 40.00% compared to the realization at $H\Delta_\tau = 10$), and 0.14 for large-tick stocks (with a percentage decrease of 35.71% compared to the realization at $H\Delta_\tau = 10$). Focusing on inter-models' dynamics, we observe that, also in this case, for small- to medium-tick stocks, performances are very similar for all the 3 evaluation metrics, except for iTransformer and LobTransformer architectures; however, differently from what observed at $H\Delta_\tau = 10$, for the iTransformer model, this observation remains true also for large-tick stocks. Comparing HLOB performances with DeepLOB ones, we observe that (i) the average gain in F1 score is equal to 0.05 for small-tick stocks (with a percentage increase of 40.00% compared to what observed at $H\Delta_\tau = 10$), 0.05 for medium-tick stocks (with a percentage increase of 60.00% compared to what observed at $H\Delta_\tau = 10$), and 0.01 for large-tick stocks (with a percentage increase of 70.00% compared to what observed at $H\Delta_\tau = 10$); (ii) the average gain in MCC is equal to 0.05 for small-tick stocks (with a percentage increase of 20.00% compared to what observed at $H\Delta_\tau = 10$), 0.02 for medium-tick stocks (with no increase compared to what observed at $H\Delta_\tau = 10$), and 0.03 for large-tick stocks (with a percentage increase of 33.33% compared to what observed at $H\Delta_\tau = 10$); (iii) the average gain in p_T is equal to 0.03 for large-tick stocks (with a percentage increase of 33.33% compared to what observed at $H\Delta_\tau = 10$), 0.03 for medium-tick stocks (with a percentage increase of 66.00% compared to what observed at $H\Delta_\tau = 10$), and 0.01 for large-tick stocks (with a percentage increase of 100.00% compared to what observed at $H\Delta_\tau = 10$).

Looking at Table 6, we notice that, at $H\Delta_\tau = 100$, the HLOB model outperforms state-of-the art (SOTA) alternatives in the 33% of cases (37% less than what happens at $H\Delta_\tau = 10$ and 27% less than what happens at $H\Delta_\tau = 50$). For small-tick stocks, HLOB is the best-performing model in 1/6 scenarios (in particular for IBM stock); in the case of CHTR, it is the second-best alternative, while in all the other cases (i.e., GOOG, GS, MCD, and NVDA), it is the third-best alternative. For medium-tick stocks, it is the third-best performing model in 3/3 scenarios (i.e., AAPL, ABBV, PM). For large-tick stocks, it is the best-performing model in 4/6 cases (i.e., BAC, CSCO, KO, ORCL), being the second-best alternative in the case of PFE and the third-best alternative in the case of VZ.

The *HLOB average F1 score* is equal to 0.32 for small-tick stocks (with a percentage decrease of 31.25% compared to the realization at $H\Delta_\tau = 10$ and a decrease of 12.50% compared to the realization at $H\Delta_\tau = 50$), 0.35 for medium-tick stocks (with a percentage decrease of 17.14% compared to the realization at $H\Delta_\tau = 10$ and a decrease of 14.28% compared to the realization at $H\Delta_\tau = 50$), and to 0.52 for large-tick stocks (with a percentage increase of 7.69% compared to the realization at $H\Delta_\tau = 10$ and a decrease of 11.54% compared to the realization at $H\Delta_\tau = 50$).

The *HLOB average MCC score* is equal to 0.05 for small-tick stocks (with a percentage decrease of 220.00% compared to the realization at $H\Delta_\tau = 10$ and a decrease of 80.00% compared to the realization at $H\Delta_\tau = 50$), 0.06 for medium-tick stocks (with a percentage decrease of 166.67% compared to the realization at $H\Delta_\tau = 10$ and a decrease of 83.33% compared to the realization at $H\Delta_\tau = 50$), and to 0.30 for large-tick stocks (with a percentage decrease of 10.00% compared to the realization at $H\Delta_\tau = 10$ and a decrease of 26.66% compared to the realization at $H\Delta_\tau = 50$).

The *HLOB average p_T score* is equal to 0.05 for small-tick stocks (with a percentage decrease of 120.00% compared to the realization at $H\Delta_\tau = 10$ and a decrease of 40.00% compared to the realization at $H\Delta_\tau = 50$), 0.07 for medium-tick stocks (with a percentage decrease of 100.00% compared to the realization at $H\Delta_\tau = 10$ and a decrease of 42.86% compared to the realization at $H\Delta_\tau = 50$), and to 0.15 for large-tick stocks (with a percentage increase of 40.00% compared to the realization at $H\Delta_\tau = 10$ and an increase of 6.67% compared to the realization at $H\Delta_\tau = 50$). Focusing on inter-models' dynamics, we observe that, consistently with what observed at $H\Delta_\tau \in \{10, 50\}$, for small- to medium-tick stocks, performances are similar for all the 3 evaluation metrics with very minor oscillations; the two main exceptions are represented by the iTransformer and LobTransformer which present considerably lower realizations. Similarly to $H\Delta_\tau = 50$, but differently from $H\Delta_\tau = 10$, this behavior persists also for large-tick stocks.

Comparing HLOB performances with DeepLOB ones, we observe that (i) the average gain in F1 score is equal to 0.006 for small-tick stocks (with a percentage decrease of 400.00% compared to what observed at $H\Delta_\tau = 10$ and a decrease of 733.33% compared to what observed at $H\Delta_\tau = 50$), 0.03 for medium-tick stocks (with a percentage increase of 33.33% compared to what observed at $H\Delta_\tau = 10$ and a decrease of 66.67% compared to what observed at $H\Delta_\tau = 50$), and 0.04 for large-tick stocks (with a percentage increase of 92.50% compared to what observed at $H\Delta_\tau = 10$ and an increase of 75.00% compared to what observed at $H\Delta_\tau = 50$); (ii) the average gain in MCC is equal to 0.04 for small-tick stocks (with no percentage increase compared to what observed at $H\Delta_\tau = 10$ and a decrease of 25.00% compared to what observed at $H\Delta_\tau = 50$), 0.02 for medium-tick stocks (with no percentage increase compared to what observed at $H\Delta_\tau \in \{10, 50\}$), and 0.04 for large-tick stocks (with a percentage increase of 100.00% compared to what observed at $H\Delta_\tau = 10$ and an increase of 33.33%

Table 6
Models' performances at $H\Delta_\tau = 100$. For each deep learning architecture we report three key metrics: (i) the F1 score; (ii) the MCC; and (iii) the p_T. For each stock, we highlight the best performing model (green), the second-best performing model (blue) and the worst performing alternative (red); a model is considered superior to the others if the sum of the 3 performance metrics is maximal.

	H100																											
	cnn1			cnn2			dla			transformer			itransformer			lobtransformer			binbtabl			binctabl			deeplob			hlob		
	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T	F1	MCC	p_T
CHTR	0.29	0.04	0.02	0.30	0.00	0.02	0.32	0.05	0.03	0.31	0.06	0.03	0.25	0.01	0.03	0.23	0.00	0.01	0.34	0.11	0.04	0.35	0.10	0.04	0.34	0.03	0.04	0.35	0.08	0.04
GOOG	0.29	0.04	0.03	0.28	0.01	0.02	0.26	0.03	0.03	0.32	0.06	0.04	0.25	0.02	0.03	0.24	0.02	0.01	0.39	0.13	0.06	0.40	0.14	0.07	0.32	0.01	0.05	0.36	0.10	0.05
GS	0.29	0.01	0.04	0.30	0.01	0.03	0.28	0.01	0.04	0.28	0.01	0.04	0.22	0.01	0.00	0.22	0.00	0.00	0.28	0.06	0.06	0.29	0.06	0.06	0.31	0.00	0.06	0.30	0.02	0.05
IBM	0.29	0.02	0.02	0.32	0.01	0.04	0.29	0.03	0.04	0.28	0.03	0.04	0.28	0.01	0.00	0.26	0.01	0.02	0.29	0.03	0.06	0.29	0.03	0.06	0.29	0.00	0.04	0.32	0.03	0.06
MCD	0.24	0.02	0.03	0.28	0.00	0.03	0.26	0.03	0.03	0.27	0.03	0.05	0.22	0.01	0.03	0.25	0.01	0.02	0.30	0.05	0.06	0.32	0.05	0.07	0.29	0.01	0.06	0.29	0.04	0.05
NVDA	0.23	0.01	0.01	0.17	0.00	0.00	0.24	0.02	0.03	0.23	0.01	0.01	0.18	0.00	0.01	0.22	0.00	0.00	0.34	0.09	0.08	0.36	0.09	0.07	0.31	0.01	0.03	0.28	0.05	0.05
AAPL	0.34	0.06	0.07	0.30	0.05	0.03	0.33	0.06	0.08	0.33	0.06	0.06	0.24	0.02	0.01	0.27	0.03	0.01	0.37	0.08	0.09	0.38	0.08	0.09	0.34	0.07	0.05	0.35	0.08	0.07
ABBV	0.26	0.04	0.03	0.32	0.02	0.03	0.33	0.05	0.08	0.31	0.05	0.06	0.22	0.03	0.03	0.24	0.01	0.02	0.35	0.07	0.08	0.36	0.08	0.08	0.31	0.03	0.05	0.34	0.06	0.06
PM	0.33	0.03	0.06	0.27	0.00	0.02	0.32	0.04	0.08	0.33	0.04	0.06	0.19	0.01	0.01	0.23	0.00	0.00	0.35	0.05	0.08	0.36	0.05	0.08	0.31	0.01	0.05	0.35	0.04	0.07
BAC	0.56	0.36	0.10	0.55	0.34	0.12	0.55	0.35	0.13	0.57	0.38	0.14	0.41	0.21	0.09	0.52	0.29	0.14	0.58	0.36	0.17	0.58	0.37	0.18	0.57	0.37	0.16	0.63	0.44	0.17
CSCO	0.52	0.28	0.16	0.52	0.29	0.15	0.50	0.27	0.18	0.52	0.29	0.15	0.38	0.14	0.12	0.47	0.23	0.12	0.49	0.27	0.16	0.50	0.27	0.15	0.52	0.28	0.14	0.52	0.30	0.18
KO	0.50	0.26	0.14	0.48	0.23	0.11	0.50	0.26	0.16	0.49	0.26	0.14	0.35	0.12	0.09	0.47	0.23	0.11	0.48	0.26	0.13	0.50	0.27	0.12	0.51	0.26	0.13	0.53	0.30	0.16
ORCL	0.48	0.23	0.14	0.46	0.21	0.12	0.45	0.22	0.16	0.47	0.24	0.16	0.30	0.07	0.08	0.48	0.24	0.16	0.48	0.22	0.17	0.48	0.22	0.17	0.48	0.25	0.16	0.49	0.26	0.17
PFE	0.50	0.26	0.13	0.53	0.29	0.12	0.48	0.26	0.13	0.48	0.27	0.12	0.38	0.12	0.09	0.49	0.27	0.10	0.52	0.27	0.16	0.50	0.27	0.14	0.50	0.28	0.11	0.50	0.28	0.13
VZ	0.45	0.19	0.12	0.41	0.17	0.12	0.42	0.19	0.13	0.45	0.20	0.13	0.28	0.07	0.07	0.36	0.14	0.08	0.47	0.21	0.16	0.47	0.21	0.17	0.33	0.12	0.10	0.47	0.22	0.11

compared to what observed at $H\Delta_\tau = 50$); (iii) the average gain in p_T is equal to 0.003 for small-tick stocks (with a percentage decrease of 566.67% compared to what observed at $H\Delta_\tau = 10$ and a decrease of 900.00% compared to what observed at $H\Delta_\tau = 50$), 0.02 for medium tick stocks (with a percentage increase of 50.00% compared to what observed at $H\Delta_\tau = 10$ and a decrease of 50.00% compared to what observed at $H\Delta_\tau = 50$), and 0.02 for large-tick stocks (with a percentage increase of 200.00% compared to what observed at $H\Delta_\tau = 10$ and an increase of 100.00% compared to what observed at $H\Delta_\tau = 50$).

In Fig. 5, we report, for each prediction horizon $H\Delta_\tau \in \{10, 50, 100\}$ and for each considered model, the distribution of p_T as a function of the total number of executed round-trip transactions (TT) (Briola et al., 2024). In this case we do not distinguish between different classes of stocks, and we use different markers to report the average models' performance values. We divide each plot into four quadrants. The *upper-left quadrant* (i.e., (I)), contains the architectures with an average TT lower than the 25% percentile (computed across the totality of models), and an average p_T greater than the 75% percentile (computed across the totality of models); intuitively, models located in this quadrant are the ones that achieve the best performances while remaining parsimonious in terms of executed transactions. The *upper-right quadrant* (i.e., (II)) contains the architectures with an average TT higher than the 25% percentile (computed across the totality of models), and an average p_T greater than the 75% percentile (computed across the totality of models); intuitively, models located in this quadrant are the ones that achieve the best performances being less parsimonious in terms of executed transactions. The *lower-left* quadrant (i.e., (III)) contains the architectures with an average TT lower than the 25% percentile (computed across the totality of models), and an average p_T lower than the 75% percentile (computed across the totality of models); intuitively, models located in this quadrant are the ones that achieve the worst performances being, however, parsimonious in terms of executed transactions. Finally, The *lower-right* quadrant (i.e., (IV)) contains the architectures with an average TT higher than the 25% percentile (computed across the totality of models), and an average p_T higher than the 75% percentile (computed across the totality of models); intuitively, models located in this quadrant are the ones that achieve the worst performances being less parsimonious in terms of executed transactions.

From this representation, it is possible to distinguish between 3 groups of models showing a consistent behavior across prediction horizons. The *first group* of models is made of BinBTabl, BinCTabl and HLOB. They are always placed in the upper-right quadrant, demonstrating to be the most effective models in correctly predicting round-trip transactions, even if not being particularly parsimonious in number of trading actions. At $H\Delta_\tau = 10$, HLOB is more effective than the other two benchmark alternatives, showing, however, yet the most pronounced attitude to perform active trading actions. This tendency disappears moving to $H\Delta_\tau \in \{50, 100\}$, where HLOB demonstrates to be slightly inferior to its benchmark alternatives. The *second group* of models is made of iTransformer and LobTransformer. They are always placed in the lower-left quadrant, demonstrating the worst performances in terms of practicability of forecasts. Finally, the *third group* is the most heterogeneous one and is made of CNN1, CNN2, DLA, Transformer, and DeepLOB architectures. Among them, DeepLOB and Transformer are the only two models which permanently remain in the same quadrant, demonstrating to be less parsimonious in terms of predicted transaction, but more accurate in terms of round-trip transactions' forecast. CNN1, CNN2, and DLA, on the other side, independently from the prediction horizon, present borderline behaviors, often placing themselves in an area between the third and the fourth quadrant.

HLOB-related results discussed previously in this Section allow us to formulate new considerations on the LOB microstructural working mechanics. As explained in Section 4, the success of the proposed architecture mainly relies on the meaningfulness of higher-order dependency structures captured by the underlying IFN. Its effectiveness is evident and persistent across different prediction horizons since it ties SOTA performances in the case of BinCTabl and BinBTabl, constantly outperforming other benchmark alternatives and, specifically, the DeepLOB model, which represents its structure-agnostic ancestor. Specifically, compared to DeepLOB, HLOB demonstrates a broad effectiveness, capturing two microstructural aspects: (i) the LOB has an underlying spatial structure that requires the modeling of higher-order and non-trivial dependency structures among volume- (and price-) levels; (ii) the emergence of dependency structures can be modeled as a function of the asset's tick size (i.e., small-, medium-, and large-tick) and their persistence varies depending on the same factor at different prediction horizons. The *first finding* can be directly derived by observing that DeepLOB, which has an architecture conceptually similar to HLOB but designed to act only on consecutive LOB's volume- and price-levels, is less effective than HLOB at all prediction horizons, remarking the necessity for modeling higher-order and deeper dependency structures. The *second finding* can be derived from the observation of HLOB's performance across prediction horizons for different classes of stocks. At $H\Delta_\tau = 10$, HLOB is superior to all the other models independently of the stocks' tick size, showing that the average structure extracted through the IFN effectively models short-term mid-price change dynamics. At $H\Delta_\tau = 50$, HLOB remains effective for medium- to large-tick stocks, where the risk (expressed via the bid–ask spread) and the LOB's actual depth (see the work by Briola et al. (2024)) are lower. This is not true for small-tick stocks, where the average structure captured by the TMFG is less robust to the LOB's changes. The same findings apply to $H\Delta_\tau = 100$, where, however, the average structure is also ineffective for medium-tick stocks. At $H\Delta_\tau \in \{50, 100\}$, HLOB is superior to DeepLOB,

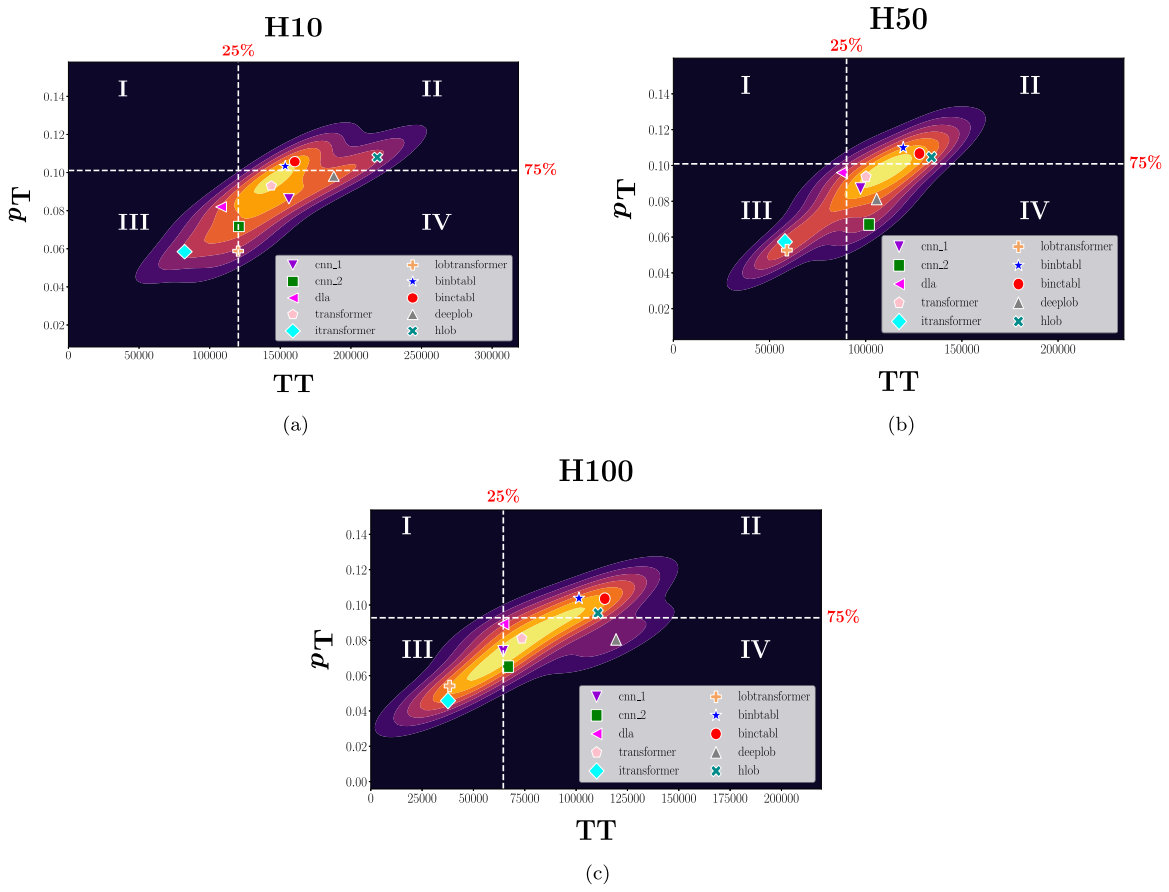

Fig. 5. Distribution of p_T (see the work by Briola et al. (2024)) as a function of the total number of executed round-trip transactions (TT) computed for each model in Table 3 at $H\Delta_\tau \in \{10, 50, 100\}$.

but slightly inferior to BinBTabl and BinCTabl. We postulate that these last two models perform better on longer horizons because they apply a dual-attention mechanism on the input's spatial and temporal dimension (the IFN behind the HLOB model only handles spatial dynamics). This means that they orchestrate a selective focus on specific elements of the input by assigning varying weights indicative of their relative significance for the given task across spatial and time LOB features, allowing the refinement of the captured non-linear relationships across time. This advantage comes with a drawback. Indeed, the high level of interpretability offered by the standard attention decreases in systems employing the dual-attention mechanism due to the inherent complexity of capturing evolving spatial dependencies over time, leading to a more dynamic and nuanced understanding of the data that might not be as easily interpreted statically. The supremacy of BinBTabl and BinCTabl models is finally annihilated in the case of large-tick stocks, which show a higher level of structure across volume levels, avoiding informational drifts that are damaging in the case of deep learning models.

5.2. Spatial distribution of information in limit order books

As a further instrument to understand the theoretical implications of empirical results obtained in Section 5.1, in Figs. 6, 7, 8, we report the average (computed across the 3-year analysis period) MI matrices computed on the training set for each of the 15 stocks under investigation (see Section 4.1). This analysis sheds light on (i) the volume levels where most of the LOB information is concentrated, and (ii) how different spatial distributions impact the model's forecasting capabilities[13]. As average matrices, the ones presented in the following Figures are not used to build the HLOB. However, they are useful in capturing the intuition behind the scenario-dependent effectiveness of the HLOB model.

Fig. 6 reports the normalized average MI matrices for *small-tick stocks*. CHTR and GOOG present similar dynamics. Their not-normalized average mutual information is 0.35 and 0.26, respectively. In the case of CHTR, we observe a weak hierarchical organization across LOB levels. The best ask and bid volume levels (i.e., $v_1^{s\in\{\text{ask, bid}\}}$) present the highest cumulative mutual informational, which smoothly decreases moving to deeper levels. The highest punctual realizations of the chosen similarity measure are generally expressed among contiguous levels on the same side of the LOB. In the case of GOOG, we notice that the decrease in the cumulative mutual information across volume levels is steeper, with a clear break after $v_4^{s\in\{\text{ask, bid}\}}$. Also in this case, the highest punctual realizations of the chosen similarity measure are generally expressed among contiguous levels on the same side of the LOB. GS presents a not-normalized average mutual information equal to 0.45. Compared to the previous two alternatives, this value turns out to be not only higher, but also differently spatially distributed. Indeed, looking at the volume levels' cumulative mutual information, we isolate three different groups: (i) $v_{\ell\in\{1,3\}}^{s\in\{\text{ask, bid}\}}$, (ii) $v_{\ell\in\{4,7\}}^{s\in\{\text{ask, bid}\}}$, and (iii) $v_{\ell\in\{8,10\}}^{s\in\{\text{ask, bid}\}}$. The first and the last group are characterized by a similar cumulative mutual information value, which, however is lower than the one of the second group.

[13] Similar attempts were performed by Libman, Ariel, Schaps, and Haber (2022) and Cont, Cucuringu, and Zhang (2023). However, their works differ from ours both in terms of the adopted methodology, granularity of analysis and results' interpretation.

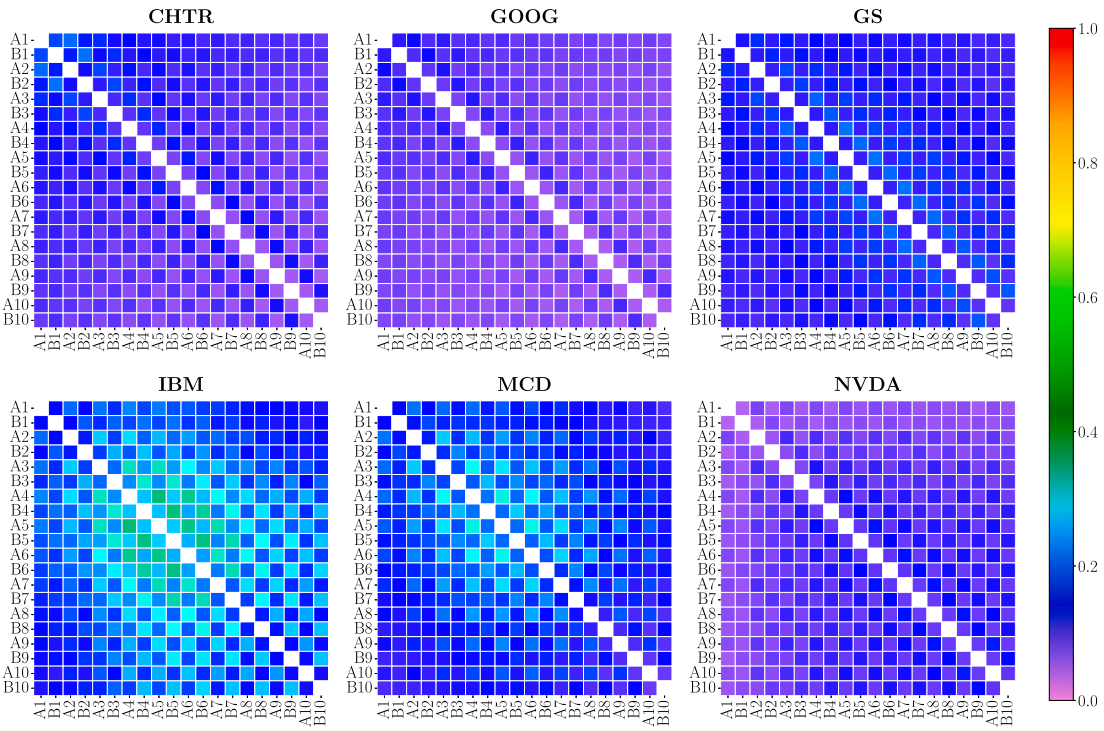

Fig. 6. Normalized (over the 15 stocks in Table 1) version of the average (computed across the 3-year analysis period) MI matrices computed on small-tick stocks (i.e., CHTR, GOOG, GS, IBM, MCD, NVDA). For the sake of readability, we renamed LOB volume levels following a mapping schema that can be summarized as follows $v_{\ell}^{\text{ask}} \rightarrow \text{A}\ell$, $v_{\ell}^{\text{bid}} \rightarrow \text{B}\ell$.

These three clusters are clearly separated, with an absence of smoothed transition. In this sense, GS presents the first signs of a hierarchical structure where the central levels of the LOB play an increasingly central role. This behavior is markedly evident in the case of IBM. This stock presents a not-normalized average mutual information equal to 0.74 (i.e., the highest among small-tick stocks), with a clear concentration towards the central levels of the LOB (i.e. $v_{\ell\in\{4,6\}}^{s\in\{\text{ask, bid}\}}$). In this case, the transition from volume levels characterized by a lower cumulative mutual information to volume levels characterized by a higher cumulative mutual information is smooth and incremental moving from top to middle volume levels and is even less evident moving from middle to deep ones (i.e. $v_{\ell\in\{7,10\}}^{s\in\{\text{ask, bid}\}}$), which are organized in a clear cluster with a medium-to-high level of cumulative mutual information. Even if visually similar to IBM, the MI matrix characterizing MCD conveys a different message. Here, the not-normalized average mutual information is equal to 0.58 and is mostly distributed across the top 8 levels of the LOB (i.e., $v_{\ell\in\{1,8\}}^{s\in\{\text{ask, bid}\}}$). In this sense, the emerging hierarchical structure is less clear compared to the one of IBM, and more similar to the one of GS. The case of NVDA, finally, is unique in the class of small-tick stocks. Here, the average mutual information is equal to 0.31 and is mainly concentrated on the deepest 6 levels of the LOB. Complementary to what observed for CHTR and GOOG, the top volume levels (i.e., $v_{1}^{s\in\{\text{ask, bid}\}}$) are characterized by the lowest cumulative mutual information, which incrementally increases moving to deeper levels of the LOB. However, also in this case, the highest punctual realizations of the chosen similarity measure are generally expressed among contiguous levels on the same side of the LOB. All the results discussed above directly derive from one of the findings in the work by Briola et al. (2024). There, the authors, following the intuition proposed by Wu et al. (2021), introduce Ξ^{Bid} and Ξ^{Ask} to measure the 'actual LOB depth' (see Table 7) on the bid and ask side of the LOB, respectively. Indeed, as described in Section 2.1, the LOB representation characterizing the data used in the current paper, suffers a lack of homogeneity in the spatial structure (since there is no assumption for adjacent price levels to be separated by fixed intervals). As a consequence, when the average $\Xi^{\{\text{Bid, Ask}\}} \gg 9.0$, as it happens for CHTR and GOOG, the computation of the average mutual information across levels is negatively affected due to the drifts that make the concept of 'level' a pure theoretical artifact with a short-term practical feedback. On the contrary, the meaningfulness of MI matrices and, consequently, the persistence of related higher-order structures across longer time horizons, increases when $\Xi^{\{\text{Bid, Ask}\}} \simeq 9.0$, with IBM providing an example of ideal environment to challenge spatially-informed deep learning models.

Fig. 7 reports the normalized average MI matrices for *medium-tick stocks*. AAPL is characterized by a not-normalized average mutual information equal to 0.41 which is mainly concentrated across $v_{\ell\in\{2,10\}}^{s\in\{\text{ask, bid}\}}$. The top volume levels of the LOB are markedly detached from the others, which, in contrast, show a strong interdependence. This behavior is far from that of ABBV and PM, which have a not-normalized average mutual information equal to 0.59 and 0.63, respectively. In both cases, the distribution of the chosen similarity metric is very similar to the one of MCD, with most of the mutual information concentrated on the top 7 volume levels of the LOB and a clear drop for the remaining 3 ones. Looking at Table 7, we notice that, in the case of AAPL, a lower average mutual information is compensated by a higher stability of the LOB, which increases the persistence of the structure extracted from the MI matrix through the IFN (described in Section 4.1). ABBV and PM, in contrast, present average Ξ^{Bid} and Ξ^{Ask} values that are more similar to the ones observed for small-tick stocks and are consequently exposed to the adverse consequences described previously in this Section.

Fig. 8 reports the normalized average MI matrices for *large-tick stocks*. The not-normalized average mutual information of BAC is equal to 1.18. It is unevenly spatially distributed across LOB levels with an evident hierarchical organization: (i) $v_{1}^{s\in\{\text{ask, bid}\}}$ express the lowest pairwise mutual information realizations; (ii) $v_{\ell\in\{2,3\}}^{s\in\{\text{ask, bid}\}}$ express an intermediate amount of pairwise mutual information; and (iii) $v_{\ell\in\{4,10\}}^{s\in\{\text{ask, bid}\}}$ contain the highest concentration of mutual information. For each of these 3 groups, it is possible to notice a smooth decrease of mutual information moving from top to deeper LOB levels. A similar dynamics can be observed for all the other stocks characterized by the same tick size, with the only difference of being able to clearly identify two

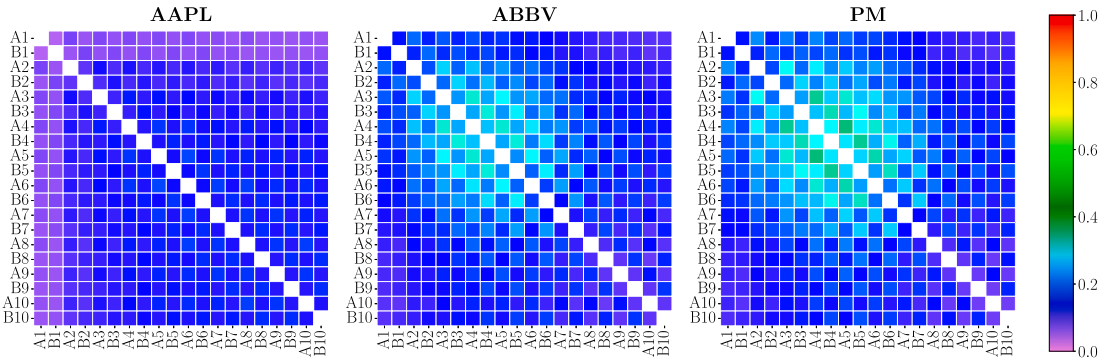

Fig. 7. Normalized (over the 15 stocks in Table 1) version of the average (computed across the 3-year analysis period) MI matrices computed on medium-tick stocks (i.e., AAPL, ABBV, PM). For the seek of readability, we renamed LOB volume levels following a mapping schema that can be summarized as follows $v_{\ell}^{\text{ask}} \rightarrow \text{A}\ell$, $v_{\ell}^{\text{bid}} \rightarrow \text{B}\ell$.

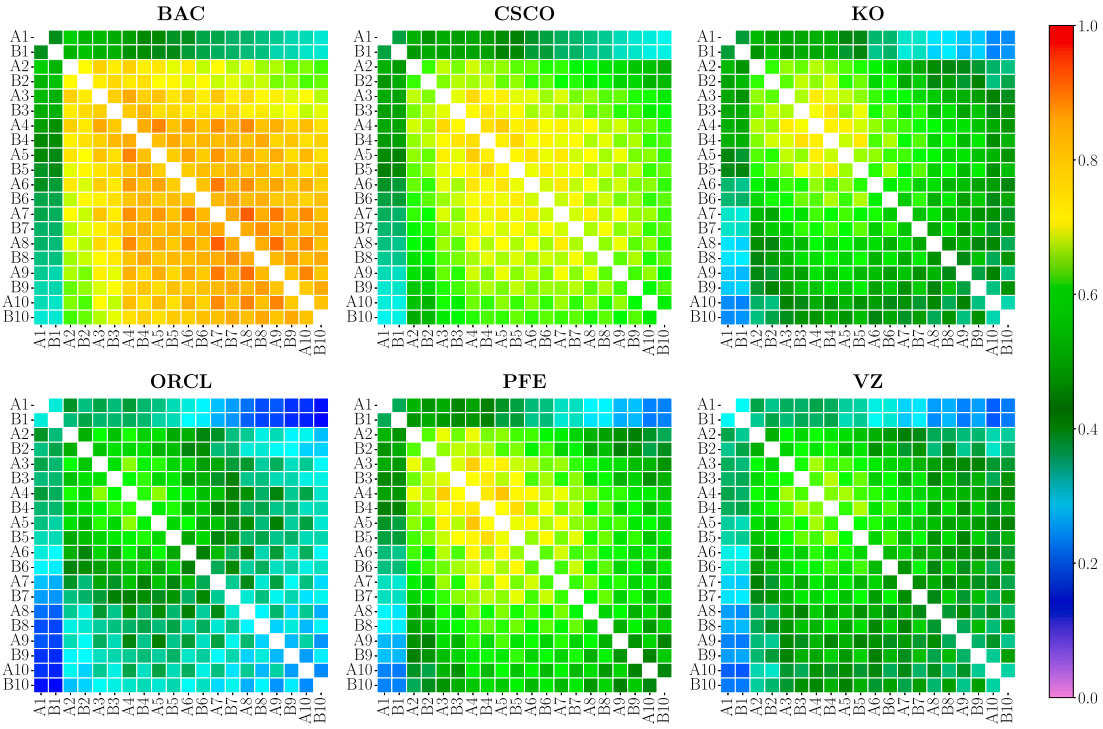

Fig. 8. Normalized (over the 15 stocks in Table 1) version of the average (computed across the 3-year analysis period) MI matrices computed on large-tick stocks (i.e., BAC, CSCO, KO, ORCL, PFE, VZ). For the seek of readability, we renamed LOB volume levels following a mapping schema that can be summarized as follows $v_{\ell}^{\text{ask}} \rightarrow \text{A}\ell$, $v_{\ell}^{\text{bid}} \rightarrow \text{B}\ell$.

(instead of three) groups of volume levels. In all the cases, $v_1^{s\in\{\text{ask, bid}\}}$ express the lowest pairwise mutual information, while $v_{\ell\in\{3,5\}}^{s\in\{\text{ask, bid}\}}$ contain the highest realizations. The not-normalized average mutual information of CSCO, KO, ORCL, PFE, and VZ is equal to 1.00, 0.83, 0.62, 0.90, and 0.74, respectively. ORCL and VZ present the lowest realizations; however, compared to small- to medium-tick stocks, they maintain a considerably higher level of structure. Looking at Table 7, we observe that all the stocks consistently exhibit the lowest realizations of Ξ^{Bid} and Ξ^{Ask}. This indicates a high level of stability in LOB structures, consequently justifying the advantage of deep learning-based models in the related forecasting tasks. Additionally, the distinct emerging structure observed across LOB levels supports their sustained effectiveness over extended prediction horizons.

To summarize, we state that (i) *small- and medium-tick stocks* generally suffer from a lack of structure in the LOB informational content, which causes a faster degradation of deep learning models' forecasting capabilities moving from closer to farther prediction horizons; (ii) while *large-tick stocks* present a more compact and meaningful structure of the LOB, guaranteeing a direct mapping between the theoretical concept of 'level' and its practical realization as an informational channel for deep learning models, which consequently has a positive effect on deep learning models' forecasting performances at both closer and farther prediction horizons.

6. Conclusion and future work

This paper introduces HLOB, a novel large-scale deep learning architecture for high-frequency Limit Order Book (LOB) mid-price changes' direction forecasting. The novelty of the model lies in the possibility to deterministically model higher-order interactions among LOB volume (and price) levels leveraging the power of a class of Information Filtering Networks (IFNs): the Triangulated Maximally Filtered Graph

Table 7
Mean and median 'actual depth' for bid and ask side of the LOB (i.e., Ξ^{Bid} and Ξ^{Ask}) for the 15 stocks of interest, in the 3-year analysis period.

	2017				2018				2019			
	Mean		Median		Mean		Median		Mean		Median	
	Ask	Bid	Ask	Bid	Ask	Bid	Ask	Bid	Ask	Bid	Ask	Bid
CHTR	53.83	50.14	34.00	35.00	71.67	68.84	55.00	54.00	44.56	45.64	35.00	36.00
GOOG	55.45	53.57	49.00	47.00	93.67	91.83	82.00	81.00	61.92	62.14	58.00	58.00
GS	14.97	15.30	13.00	13.00	19.07	20.33	15.00	16.00	13.98	14.12	13.00	13.00
IBM	10.29	10.47	9.00	10.00	12.47	12.70	11.00	11.00	9.93	9.93	9.00	9.00
MCD	126.97	9.95	9.00	9.00	12.32	12.73	10.00	11.00	13.35	13.59	12.00	12.00
NVDA	10.79	10.73	9.00	9.00	16.30	16.12	15.00	15.00	10.87	10.88	10.00	10.00
AAPL	9.02	9.02	9.00	9.00	9.52	9.54	9.00	9.00	9.13	9.14	9.00	9.00
ABBV	10.95	11.13	9.00	9.00	23.27	19.79	11.00	11.00	9.74	9.69	9.00	9.00
PM	10.09	10.07	9.00	9.00	13.64	13.60	11.00	11.00	10.22	10.20	9.00	9.00
BAC	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00
CSCO	9.00	9.00	9.00	9.00	9.01	9.01	9.00	9.00	9.00	9.00	9.00	9.00
KO	9.14	9.19	9.00	9.00	9.04	9.04	9.00	9.00	9.01	9.01	9.00	9.00
ORCL	9.11	9.11	9.00	9.00	9.05	9.06	9.00	9.00	9.01	9.01	9.00	9.00
PFE	9.20	9.21	9.00	9.00	9.03	9.03	9.00	9.00	9.01	9.01	9.00	9.00
VZ	9.09	9.09	9.00	9.00	9.07	9.09	9.00	9.00	9.01	9.01	9.00	9.00

(TMFG). Its computation exploits the pairwise mutual information across volume levels, and its structure only retains statistically relevant dependencies, pruning the weakest ones. The informational content of the emerging homological priors (i.e., tetrahedra, triangles, and edges) is then mapped as input to the class of Homological Convolutional Neural Networks (HCNNs), and processed to forecast the direction of high-frequency mid-price changes for 15 stocks belonging to 3 different classes (i.e., small-, medium-, and large-tick stocks) over a 3-year analysis period (i.e., 2017, 2018, and 2019). This class of neural networks naturally models the spatial dimension of the LOB and is modified here to handle long-term temporal dependencies through the introduction of the long short-term memory (LSTM) module. This modification sets off the transition from a simple HCNN to an HLOB model. The development of this architecture is backed by the hypothesis that a more structured architectural grasp of the LOB's spatial dependency structures would enhance a model's predictive precision. To test this hypothesis, we test our architecture against 9 SOTA models, demonstrating not only the supremacy of our architecture in specific scenarios but also using the empirical results to prove some theoretical conjectures on the microstructural mechanics of the LOBs. Specifically, we obtain 3 main findings:

- We demonstrate that the LOB has an underlying spatial structure that requires modeling higher-order and non-trivial dependency structures among volume (and price) levels, making the description of interactions among consecutive levels (used for instance in DeepLOB (Zhang, Zohren, & Roberts, 2019)) a sub-optimal solution. This result is achieved by leveraging the Triangulated Maximally Filtered Graph (TMFG) as a tool for undirected graphical modeling, allowing for a comprehensive investigation of both local and non-local interactions within the LOB. Through a likelihood maximization framework, the chosen model captures intricate dependency structures that encompass both immediate and deeper spatial relationships among volume and price levels.
- We show that the emergence of dependency structures can be modeled as a function of the asset's tick size; different structures emerge for different types of assets, and the ones characterized by a clear hierarchical structure (i.e., large-tick stocks) have more chances to be correctly forecast by deep learning models.
- We demonstrate that the persistence of the informational content that can be captured through a modeling exercise on the LOB spatial dependency structure varies at different prediction horizons. Indeed, when the LOB structure is sparse and subject to informational drifts (i.e., small- and medium-tick stocks), the concept of 'level' becomes a purely theoretical artifact with a limited-in-time realization in practical scenarios. In this case, a deep learning model built on the average mutual information across volume levels is also exposed to the adverse impact of outliers (i.e., informational drifts) being effective only at short-term prediction horizons, where the likelihood of informational drifts is lower.

HLOB represents a step forward in building microstructurally-informed models for the prediction of the direction of high-frequency mid-price changes. The overall performance of the proposed architecture is commendable. It marks a significant advancement in the field of microstructural modeling, providing practitioners and academics with a powerful tool that combines the power of deep learning with a nuanced understanding of LOB mechanics, facilitating better decision-making in high-frequency environments. From its comparison with alternative SOTA models, some limitations emerge. Specifically, applying a dual-attention mechanism on the input data spatial and temporal dimensions guarantees, at the price of a reduced interpretability, enhanced performances, allowing for a refinement of the captured non-linear relationships over time and offering an edge over HLOB, which primarily handles spatial dynamics.

There are several avenues for further development of the HLOB model. As a future research work, (i) we should reason about more refined ways to compute the similarity matrices at the core of the proposed architecture and, as a consequence of this, (ii) we should think about the possibility of modifying the HLOB to incorporate temporally-evolving IFNs capturing the evolving complexity of the LOB. Further, from the market microstructure perspective, it would be useful to repeat our experiments on data from exchanges other than the NASDAQ one in order to study how different or temporally evolving tick sizes impact our observations. This research marks an initial step towards developing microstructurally informed models that can adapt to the complexities of high-frequency market phase transitions. Future work will build on these foundations to enhance models' adaptability and accuracy, paving the way for more robust and practical applications in financial markets forecasting.

CRediT authorship contribution statement

Antonio Briola: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing. **Silvia Bartolucci:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing – review & editing. **Tomaso Aste:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author, A.B., acknowledges his PhD scholarship's founder for the financial support. The author, T.A., acknowledges the financial support from ESRC, United Kingdom (ES/K002309/1), EPSRC, United Kingdom (EP/P031730/1) and EC (H2020-ICT-2018-2 825215).

Data availability

The authors do not have permission to share data.

References

Aste, T. (2022). Topological regularization with information filtering networks. *Information Sciences*, *608*, 655–669.

Aste, T., & Di Matteo, T. (2006). Dynamical networks from correlations. *Physica A. Statistical Mechanics and its Applications*, *370*(1), 156–161.

Aste, T., Di Matteo, T., & Hyde, S. (2005). Complex networks on hyperbolic surfaces. *Physica A. Statistical Mechanics and its Applications*, *346*(1–2), 20–26.

Barfuss, W., Massara, G. P., Di Matteo, T., & Aste, T. (2016). Parsimonious modeling with information filtering networks. *Physical Review E*, *94*(6), Article 062306.

Bouchaud, J.-P., Bonart, J., Donier, J., & Gould, M. (2018). *Trades, quotes and prices: financial markets under the microscope*. Cambridge University Press.

Bouchaud, J.-P., Farmer, J. D., & Lillo, F. (2009). How markets slowly digest changes in supply and demand. In *Handbook of financial markets: dynamics and evolution* (pp. 57–160). Elsevier.

Briola, A., & Aste, T. (2022). Dependency structures in cryptocurrency market from high to low frequency. *Entropy*, *24*(11), 1548.

Briola, A., Bartolucci, S., & Aste, T. (2024). Deep limit order book forecasting. arXiv preprint arXiv:2403.09267.

Briola, A., Turiel, J., & Aste, T. (2020). Deep learning modeling of limit order book: A comparative perspective. arXiv preprint arXiv:2007.07319.

Briola, A., Turiel, J., Marcaccioli, R., Cauderan, A., & Aste, T. (2021). Deep reinforcement learning for active high frequency trading. arXiv preprint arXiv:2101.07107.

Briola, A., Wang, Y., Bartolucci, S., & Aste, T. (2023). Homological convolutional neural networks. arXiv preprint arXiv:2308.13816.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

companiesmarketcap. com (2024). Companies market cap. https://companiesmarketcap.com. (Accessed 24 January 2024).

Cont, R., Cucuringu, M., & Zhang, C. (2023). Cross-impact of order flow imbalance in equity markets. *Quantitative Finance*, *23*(10), 1373–1393.

Farmer, J. D., & Skouras, S. (2013). An ecological perspective on the future of computer trading. *Quantitative Finance*, *13*(3), 325–346.

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, *28*(5–6), 367–374.

Guo, Y., & Chen, X. (2023). Forecasting the mid-price movements with high-frequency LOB: A dual-stage temporal attention-based deep learning architecture. *Arabian Journal for Science and Engineering*, *48*(8), 9597–9618.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Isichenko, M. (2021). *Quantitative portfolio management: The art and science of statistical arbitrage*. John Wiley & Sons.

Jain, K., Firoozye, N., Kochems, J., & Treleaven, P. (2024a). Limit order book dynamics and order size modelling using compound hawkes process. *Finance Research Letters*, *69*, Article 106157.

Jain, K., Firoozye, N., Kochems, J., & Treleaven, P. (2024b). Limit order book simulations: A review. arXiv preprint arXiv:2402.17359.

Jain, K., Muzy, J.-F., Kochems, J., & Bacry, E. (2024). No tick-size too small: A general method for modelling small tick limit order books. arXiv preprint arXiv:2410.08744.

Karpathy (2024). nanoGPT. https://github.com/karpathy/nanoGPT/tree/master. (Accessed 12 January 2024).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kisiel, D., & Gorse, D. (2022). Axial-LOB: High-frequency trading with axial attention. In *2022 IEEE symposium series on computational intelligence* (pp. 1327–1333). IEEE.

Kolm, P. N., Turiel, J., & Westray, N. (2023). Deep order flow imbalance: Extracting alpha at multiple horizons from the limit order book. *Mathematical Finance*, *33*(4), 1044–1081.

Kolm, P. N., & Westray, N. (2024). Improving deep learning of alpha term structures from the order book. Available at SSRN.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lehalle, C.-A., & Laruelle, S. (2018). *Market microstructure in practice*. World Scientific.

Libman, D., Ariel, G., Schaps, M., & Haber, S. (2022). Mutual information between order book layers. *Entropy*, *24*(3), 343.

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., et al. (2023). Itransformer: Inverted transformers are effective for time series forecasting. arXiv preprint arXiv:2310.06625.

LOBSTER Data (2023). What is LOBSTER? https://lobsterdata.com/info/WhatIsLOBSTER.php. (Accessed 26 December 2023).

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Lucchese, L., Pakkanen, M., & Veraart, A. (2022). The short-term predictability of returns in order book markets: a deep learning perspective. arXiv preprint arXiv:2211.13777.

Mantegna, R. N. (1999). Hierarchical structure in financial markets. *The European Physical Journal B*, *11*(1), 193–197.

Massara, G. P., Di Matteo, T., & Aste, T. (2016). Network filtering for big data: Triangulated maximally filtered graph. *Journal of Complex Networks*, *5*(2), 161–178.

Massara, G. P., Di Matteo, T., & Aste, T. (2017). Network filtering for big data: Triangulated maximally filtered graph. *Journal of Complex Networks*, *5*(2), 161–178.

NASDAQ (2023). NASDAQ stock screener. https://www.nasdaq.com/market-activity/stocks/screener. (Accessed 26 December 2023).

Ntakaris, A., Magris, M., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2018). Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting*, *37*(8), 852–866.

O'Hara, S., & Draper, B. A. (2011). Introduction to the bag of features paradigm for image classification and retrieval. arXiv preprint arXiv:1101.3354.

Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2020). Temporal logistic neural bag-of-features for financial time series forecasting leveraging limit order book data. *Pattern Recognition Letters*, *136*, 183–189.

Passalis, N., Tsantekidis, A., Tefas, A., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2017). Time-series classification using neural bag-of-features. In *2017 25th European signal processing conference* (pp. 301–305). IEEE.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Prata, M., Masi, G., Berti, L., Arrigoni, V., Coletta, A., Cannistraci, I., et al. (2023). LOB-based deep learning models for stock price trend prediction: A benchmark study. arXiv preprint arXiv:2308.01915.

Shabani, M., Tran, D. T., Kanniainen, J., & Iosifidis, A. (2023). Augmented bilinear network for incremental multi-stock time-series classification. *Pattern Recognition*, *141*, Article 109604.

Shabani, M., Tran, D. T., Magris, M., Kanniainen, J., & Iosifidis, A. (2022). Multi-head temporal attention-augmented bilinear network for financial time series prediction. In *2022 30th European signal processing conference* (pp. 1487–1491). IEEE.

Sirignano, J. A. (2019). Deep learning for limit order books. *Quantitative Finance*, *19*(4), 549–570.

Sirignano, J., & Cont, R. (2021). Universal features of price formation in financial markets: perspectives from deep learning. In *Machine learning and AI in finance* (pp. 5–15). Routledge.

Tran, D. T., Iosifidis, A., Kanniainen, J., & Gabbouj, M. (2018). Temporal attention-augmented bilinear network for financial time-series data analysis. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(5), 1407–1418.

Tran, D. T., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2021). Data normalization for bilinear structures in high-frequency financial time-series. In *2020 25th international conference on pattern recognition* (pp. 7287–7292). IEEE.

Tran, D. T., Passalis, N., Tefas, A., Gabbouj, M., & Iosifidis, A. (2022). Attention-based neural bag-of-features learning for sequence data. *IEEE Access*, *10*, 45542–45552.

Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2017a). Forecasting stock prices from the limit order book using convolutional neural networks. In *2017 IEEE 19th conference on business informatics, vol. 1* (pp. 7–12). IEEE.

Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2017b). Using deep learning to detect price change indications in financial markets. In *2017 25th European signal processing conference* (pp. 2511–2515). IEEE.

Tsantekidis, A., Passalis, N., Tefas, A., Kanniainen, J., Gabbouj, M., & Iosifidis, A. (2020). Using deep learning for price prediction by exploiting stationary limit order book features. *Applied Soft Computing*, *93*, Article 106401.

Tumminello, M., Aste, T., Di Matteo, T., & Mantegna, R. N. (2005). A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences*, *102*(30), 10421–10426.

Tumminello, M., Di Matteo, T., Aste, T., & Mantegna, R. N. (2007). Correlation based networks of equity returns sampled at different time horizons. *The European Physical Journal B*, *55*, 209–217.

UCL CS HPC Cluster (2023). UCL CS HPC cluster. https://hpc.cs.ucl.ac.uk. (Accessed 16 June 2023).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Wallbridge, J. (2020). Transformers for limit order books. arXiv preprint arXiv:2003.00130.

Wang, Y., Briola, A., & Aste, T. (2023). Homological neural networks: A sparse architecture for multivariate complexity. arXiv:2306.15337.

West, D. B., et al. (2001). *Introduction to graph theory*: *vol. 2*, Prentice hall Upper Saddle River.

Wu, Y., Mahfouz, M., Magazzeni, D., & Veloso, M. (2021). Towardss robust representation of limit orders books for deep learning models. arXiv preprint arXiv:2110.05479.

Zhang, Z., Lim, B., & Zohren, S. (2021). Deep learning for market by order data. *Applied Mathematical Finance*, *28*(1), 79–95.

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, *52*(1), 1–38.

Zhang, Z., Zohren, S., & Roberts, S. (2019). Deeplob: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing*, *67*(11), 3001–3012.