SR-Traffic: Discovering Macroscopic Traffic Flow Models with Symbolic Regression

Simone Manti
Aarhus University
smanti@mpe.au.dk

Saeed Mohammadian
The University of Queensland
s.mohammadian@uq.edu.au

Martin Treiber
Technische Universität Dresden
martin.treiber@tu-dresden.de

Alessandro Lucantonio
Aarhus University
a.lucantonio@mpe.au.dk

Abstract

Traffic flows are complex systems that can be studied from a macroscopic perspective. In particular, first-order models are tractable but oversimplified, while higher-order models capture richer dynamics at the cost of complexity. Here, we introduce *SR-Traffic*, a data-driven, physics-informed framework that uses symbolic regression to learn effective phenomenological relations directly from experimental data while embedding them into an efficient, first-order PDE formulation. Our approach balances accuracy and interpretability, ensures physical consistency, and shows good generalization, overcoming the limitations of purely data-driven models. Overall, our findings could support the design of digital solutions for improving mobility systems.

1 Introduction

Traffic flow exhibits complex phenomena such as instabilities, stop-and-go waves, and hysteresis [1, 2]. Decades of mathematical modeling have been aimed at reproducing these behaviors to support applications such as traffic control [1, 3]. Traffic models are either microscopic, simulating individual vehicle interactions [4], or macroscopic, describing aggregate quantities like density and speed [5]; this paper focuses on the latter. Macroscopic models, favored for control applications due to their simplicity, lower computational cost, and ease of calibration [3], treat traffic as a continuum governed by hyperbolic partial differential equations (PDEs) [1, 2] and can be broadly divided into first and second-order models.

First-order models rely on the continuity equation and a deterministic flow–density relation, *i.e.* the fundamental diagram (FD) [1, 6]. The seminal Lighthill-Whitham-Richards (LWR) model [7, 8] and its discretization, the Cell Transmission Model (CTM) [9], are widely used for simulation and control. Thanks to their efficiency and tractability, first-order models have been extended to include multiclass traffic [10], stochastic FDs [11], and non-local effects [12]. However, these models still fail to reproduce some key observed features, such as the wide scattering in flow–density data, which limits their performance in applications like travel-time estimation [2]. Second-order models extend first-order formulations with an acceleration equation, yielding systems derived from car-following [13] or gas-kinetic theory [1]. These models capture instabilities, stop-and-go waves, hysteresis, and flow–density scattering [1, 5], and can incorporate human factors [14], but are more complex, sometimes non-conservative [15], and require advanced numerical schemes [16, 17].

This paper proposes a novel data-driven, physics-informed macroscopic framework, named *SR-Traffic*, that combines the tractability of first-order models with the descriptive power of higher-order ones.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

We adopt a first-order continuum formulation and use symbolic regression [18, 19, 20] to learn the FD from experimental data. The availability of high-resolution datasets [21, 22, 23] has motivated the use of machine learning for traffic forecasting [24, 25] and state estimation [26, 27, 28]. However, purely data-driven approaches are black-box and often struggle at generalization (forecasting). In contrast, *SR-Traffic* improves in accuracy with respect to standard first-order models and produces models that generalize well, while maintaining interpretability, and physical consistency.

2 Methods

First order traffic flow models In macroscopic traffic flow, the continuity equation enforces the conservation of the number of vehicles:

$$\frac{\partial \rho}{\partial t} + \frac{\partial Q}{\partial x} = 0, \tag{1}$$

where ρ is traffic density and $Q = \rho v$ the flow. Lighthill–Whitham [7] and Richards [8] assumed local equilibrium:

$$Q(x,t) = Q_e(\rho(x,t)) := \rho(x,t)V(\rho(x,t)), \tag{2}$$

where Q_e is the *fundamental diagram* (FD), usually fitted from data. Some speed-density relations are listed in Table 1, with the triangular model noted for its efficiency and robustness [1]. Greenshields and Greenberg models, though not suitable to fit real-world data, are included due to their frequent use in mathematical traffic models.

Symbolic regression of fundamental diagrams Symbolic regression (SR) [18] searches for a mathematical expression that best fits a given dataset, by combining a set of operators and variables. To tackle SR, we use genetic programming (GP) [19, 20, 32], an evolutionary method where candidate expressions represented as trees are evolved through selection, crossover, mutation, and survival guided by a *fitness* measure. Here, we adopt SR to discover FDs from traffic data, given as density and velocity fields over a (x,t) domain discretized on a uniform grid $(i\Delta_x, j\Delta_t)_{i,j}$, with $i = 0, \dots, N_x - 1$ and $j = 0, \dots, N_t - 1$. We represent V in (2) as the product of a triangular (or other well-performing) speed-density and a discrete functional of ρ learned via SR, *i.e.*, $V(\rho) = V_{\text{triang}}(\rho)V_{\text{SR}}[\rho]$, and subject to the constraint $V'(\rho) \leq 0$. The triangular FD is chosen as a backbone since it is the most efficient in simulations ([1], Section 8.5). Preliminary exploratory experiments shows that the choice of the backbone FD is not critical; in other words, *SR-Traffic* consistently discovers the same models when initialized with different well-performing backbone FDs (*e.g.*, Weidmann).

Table 1: Some speed-density relations $V(\rho)$. Here, $\mathbb{1}_\Omega$ is the indicator function of the set Ω. In the following, $V_0, \rho_{\max}, \lambda_W, T, \rho_C$ are parameters to calibrate depening on the dataset.

Model	$V(\rho)$
Greenshields [29]	$V_0(1 - \rho/\rho_{\max})$
Greenberg [30]	$V_0 \log(\rho_{\max}/\rho)$
Weidmann [31]	$V_0[1 - \exp(-\lambda_W(1/\rho - 1/\rho_{\max}))]$
Triangular [9]	$V_0\mathbb{1}_{\{\rho \leq \rho_C\}} + 1/T(1/\rho - 1/\rho_{\max})\mathbb{1}_{\{\rho_C < \rho \leq \rho_{\max}\}}$

Following [33], we combine SR with discrete exterior calculus (DEC) [34, 35], a naturally discrete mathematical theory providing a set of primitives for $V_{\text{SR}}[\rho]$ that preserves geometric structure and produces physically consistent models. From the DEC viewpoint, density ρ and velocity v can be represented as a dual and a primal 0-*cochains* (discrete forms), *i.e.* collections of values sampled at the centers and the interfaces, respectively, of the space interval. The primitive set includes basic DEC operators, such as the Hodge star, the coboundary, and the codifferential δ, along with parabolic, upwind and downwind operators, *i.e.*, for a primal 0-cochain c:

$$(\flat_{par} c)_i = \frac{c_i + c_{i+1}}{2}\Delta_x, \quad (\flat_{upw} c)_i = c_i \Delta_x, \quad (\flat_{dow} c)_i = c_{i+1}\Delta_x. \tag{3}$$

Convolution operators $f *_k g := \int_0^{k\Delta_x} f(x+y)g(y)dy$, $k = 1, 2, 3$, are also included. The primitive set also includes some standard algebraic operators, precisely addition, subtraction, multiplication, division (protected), square, square root and exponential. The terminal set consists of ρ, the constant $\mathbb{1}$ primal 0-cochain and tunable scalar coefficients.

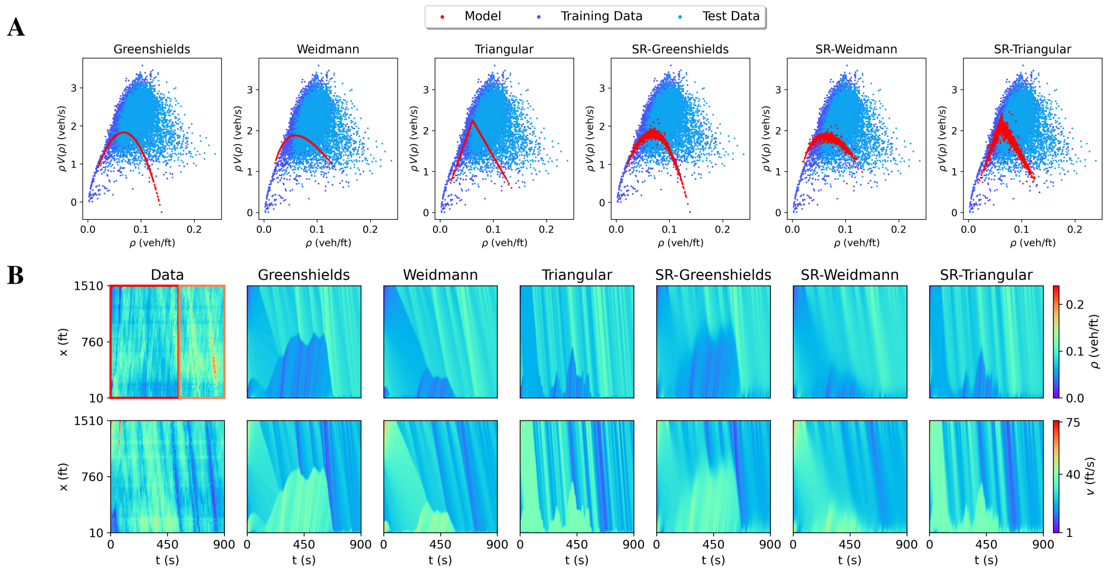

Figure 1: Traffic forecasting: (A) fundamental diagrams; (B) density and velocity fields (the red and orange rectangles correspond to the training and test regions, respectively).

Table 2: rRMSEs for density and velocity (training/test) for the traffic forecasting task.

Model	E_ρ^{tr}	E_v^{tr}	E_ρ^{ts}	E_v^{ts}	Avg Rank
Greenshields	0.299 (6)	0.256 (4)	0.289 (6)	0.304 (6)	5.50
SR-Greenshields	0.296 (5)	0.234 (2)	0.287 (5)	0.285 (5)	4.25
Weidmann	0.264 (3)	0.246 (3)	0.260 (3)	0.279 (3)	3.00
SR-Weidmann	**0.249 (1)**	**0.226 (1)**	0.250 (2)	0.261 (2)	**1.50**
Triangular	0.264 (4)	0.286 (6)	0.264 (4)	0.285 (4)	4.50
SR-Triangular	0.251 (2)	0.272 (5)	**0.248 (1)**	**0.258 (1)**	2.25

Model training consists of finding the FD that minimizes the fitness $F(\rho, v) := \frac{1}{2}(\text{rMSE}(\rho) + \text{rMSE}(v))$, where $\text{rMSE}(u) := \sum_{i=1}^{N_x} \sum_{j=1}^{N_t} |\bar{u}(x_i, t_j) - u(x_i, t_j)|^2 / \sum_{i=1}^{N_x} \sum_{j=1}^{N_t} |u(x_i, t_j)|^2$, with $\bar{u}$ and u the model predictions and the traffic data, respectively. For each candidate FD, predictions are computed by numerically solving (1)–(2) using the Godunov method [36]. The tunable constants of each FD are calibrated to minimize F, while those of V_{triang} are calibrated before running SR by minimizing the same function.

3 Results

We evaluate *SR-Traffic* on two tasks: traffic forecasting and state reconstruction. Performance is measured by relative root mean squared errors (rRMSEs) on density $E_\rho = \text{rRMSE}(\rho)$ and velocity $E_v = \text{rRMSE}(v)$, with $\text{rRMSE}(u) = \sqrt{\text{rMSE}(u)}$. We show that although V_{SR} is derived using the triangular FD as an ansatz, it remains effective, after recalibration, when applied to the other FDs, preserving the same behavior.

Experiments use the `Flex` SR library [33] and the code to reproduce them is available at https://github.com/cpml-au/SR-Traffic.

Dataset The I-80 dataset from the NGSIM program [21] contains high-resolution vehicle trajectories over a 500 m stretch of eastbound I-80 in Emeryville, CA. Data were collected on April 13, 2005, from 4:00–4:15 PM using video cameras and processed to extract positions, velocities, and accelerations. We reconstruct macroscopic density and velocity via Edie's method [37, 24].

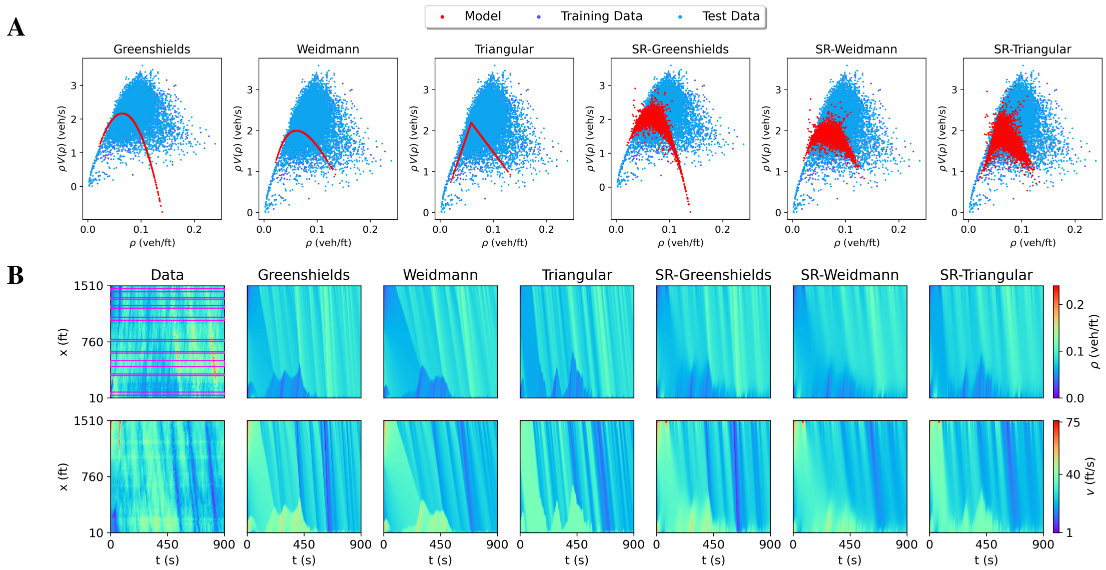

Figure 2: Traffic state reconstruction: (A) fundamental diagrams; (B) density and velocity fields (the training data are enclosed within purple rectangles).

Traffic forecasting We use the first 60% of the time interval for training and the rest for test, *i.e.* predicting the future traffic state, as shown in Figure 1B (top left). The *SR-Traffic* algorithm discovers the following model:

$$V_{\text{SR}}[\rho] = \mathbb{1} + (\delta\flat_{\text{upw}} \exp(c_1\rho)) *_1 \exp(c_2\rho), \quad \text{where } (\delta\,\flat_{\text{upw}} f)_i = \frac{f_{i+1} - f_i}{\Delta_x} \tag{4}$$

with c_1 and c_2 constants to be calibrated. We note that the SR term allows to better capture the scattering in the flow-density data (Figure 1A). Interestingly, both the density gradient (approximated via the upwind term) and the convolution may be physically interpreted as the drivers' look-ahead behavior [38]. We note that, across all the metrics considered, the models extended with the SR term outperform their corresponding base counterparts (Figure 1B and Table 2).

Traffic state reconstruction This task involves reconstructing the traffic dynamics based on the observations at a limited number of sensors. Precisely, we split the data using a random 20% of the space interval for training and the rest for test, as shown in Figure 2B (top left). The *SR-Traffic* algorithm discovers the following term:

$$V_{\text{SR}}[\rho] = \exp[(\delta\flat_{\text{dow}}\rho) *_3 (c\rho)], \quad \text{where } (\delta\,\flat_{\text{dow}} f)_i = \frac{f_i - f_{i-1}}{\Delta_x}, \tag{5}$$

where c is a constant to be calibrated. The structure of this term, which involves the exponential of a convolution, resembles previously reported models [39]. Similarly to the model in (4) found for traffic forecasting, this model better captures the scattering of the flow-density data (Figure 2A) and improves the reconstruction accuracy compared to the baselines (Figure 2B and Table 3).

4 Conclusions

We have introduced *SR-Traffic*, a symbolic regression framework for discovering fundamental diagrams of first-order, macroscopic traffic flow models from traffic data. *SR-Traffic* augments existing FDs with a multiplicative term, preserving their structure while better capturing the scattering that is typical of the flow-density data. The resulting models are interpretable and have been validated on the I-80 highway data from NGSIM for traffic forecasting and state reconstruction, where they consistently showed improved performance over standard FDs. Future work will extend *SR-Traffic* to discover more general FDs and generalize to second-order traffic models. We believe that our results may contribute to the development of digital tools for smart mobility platforms.

Table 3: rRMSEs for density and velocity (training/test) for the traffic state reconstruction task.

Model	E_ρ^{tr}	E_v^{tr}	E_ρ^{ts}	E_v^{ts}	Avg Rank
Greenshields	0.238 (4)	0.234 (4)	0.253 (4)	0.266 (4)	4.00
SR-Greenshields	0.231 (2)	0.230 (2)	0.243 (2)	0.257 (3)	2.25
Weidmann	0.242 (5)	0.231 (3)	0.258 (5)	0.252 (2)	3.75
SR-Weidmann	**0.230 (1)**	**0.215 (1)**	**0.242 (1)**	**0.234 (1)**	**1.00**
Triangular	0.250 (6)	0.251 (6)	0.265 (6)	0.285 (6)	6.00
SR-Triangular	0.233 (3)	0.235 (5)	0.246 (3)	0.266 (5)	4.00

Acknowledgments

The work of Simone Manti and Alessandro Lucantonio is supported by the European Union (European Research Council (ERC), ALPS, 101039481). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the ERC Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Computational resources have been partially provided by DeiC National HPC (DeiC-AU-N1-2023030).

References

[1] M. Treiber and A. Kesting. *Traffic Flow Dynamics*, volume 1. Springer, 2013.

[2] S. Mohammadian, Z. Zheng, M. M. Haque, and A. Bhaskar. Performance of continuum models for realworld traffic flows: Comprehensive benchmarking. *Transportation Research Part B: Methodological*, 147:132–167, 2021.

[3] M. Papageorgiou, C. Diakaki, V. Dinopoulou, A. Kotsialos, and Y. Wang. Review of Road Traffic Control Strategies. *Proceedings of the IEEE*, 91(12):2043–2067, 2003.

[4] M. Saifuzzaman and Z. Zheng. Incorporating human-factors in car-following models: A review of recent developments and research needs. *Transportation Research Part C: Emerging Technologies*, 48:379–403, 2014.

[5] S. Mohammadian, Z. Zheng, M. M. Haque, and A. Bhaskar. Continuum modeling of freeway traffic flows: State-of-the-art, challenges and future directions in the era of connected and automated vehicles. *Communications in Transportation Research*, 3:100107, 2023.

[6] M. J. Cassidy. Bivariate relations in nearly stationary highway traffic. *Transportation Research Part B: Methodological*, 32(1):49–59, 1998.

[7] M. J. Lighthill and G. B. Whitham. On kinematic waves ii. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 229(1178):317–345, 1955.

[8] P. I. Richards. Shock Waves on the Highway. *Operations Research*, 4(1):42–51, 1956.

[9] C. F. Daganzo. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation research part B: Methodological*, 28(4):269–287, 1994.

[10] Z. Qian, J. Li, X. Li, H. M. Zhang, and H. Wang. Modeling heterogeneous traffic flow: A pragmatic approach. *Transportation Research Part B: Methodological*, 99(3):183–204, 2017.

[11] D. Ngoduy. Multiclass first-order traffic model using stochastic fundamental diagrams. *Transportmetrica*, 7(2):111–125, 2011.

[12] F. A. Chiarello. An overview of non-local traffic flow models. *Mathematical Descriptions of Traffic Flow: Micro, Macro and Kinetic Models*, pages 79–91, 2020.

[13] W.-L. Jin. On the equivalence between continuum and car-following models of traffic flow. *Transportation Research Part B: Methodological*, 93:543–559, 2016.

[14] S. Mohammadian, Z. Zheng, M. Haque, and A. Bhaskar. NET-RAT: Non-equilibrium traffic model based on risk allostasis theory. *Transportation Research Part A: Policy and Practice*, 174:103731, 2023.

[15] S. Mohammadian, A. M. Moghaddam, and A. Sahaf. On the performance of HLL, HLLC, and Rusanov solvers for hyperbolic traffic models. *Computers & Fluids*, 231:105161, 2021.

[16] S. Mohammadian, Z. Zheng, S. Chu, and A. Kurganov. Numerical study of the non-conservative NET-RAT traffic flow model by path-conservative central-upwind schemes. *Computers & Mathematics with Applications*, 179:212–228, 2025.

[17] S. Chu, A. Kurganov, S. Mohammadian, and Z. Zheng. Central-upwind scheme for the phase-transition traffic flow model. *Journal of Computational Physics*, 539:114241, 2025.

[18] G. Kronberger, B. Burlacu, M. Kommenda, S. M. Winkler, and M. Affenzeller. *Symbolic Regression*. CRC Press, 2024.

[19] J. R. Koza. *On the Programming of Computers by Means of Natural Selection*. MIT press, 1992.

[20] J. R. Koza. Genetic programming as a means for programming computers by natural selection. *Statistics and computing*, 4:87–112, 1994.

[21] V. Kovvali, V. Alexiadis, and L. Zhang. Video-Based Vehicle Trajectory Data Collection. In *Transportation Research Board 86th Annual Meeting*. Citeseer, 2006.

[22] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein. The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2118–2125. IEEE, 2018.

[23] T. Moers, L. Vater, R. Krajewski, J. Bock, A. Zlocki, and L. Eckstein. The exiD Dataset: A Real-World Trajectory Dataset of Highly Interactive Highway Scenarios in Germany. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 958–964. IEEE, 2022.

[24] A. M. Avila and I. Mezić. Data-driven analysis and forecasting of highway traffic dynamics. *Nature communications*, 11(1):2090, 2020.

[25] Y. Li, R. Yu, C. Shahabi, and Y. Liu. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *International Conference on Learning Representations*, 2018.

[26] A. J. Huang and S. Agarwal. Physics-Informed Deep Learning for Traffic State Estimation: Illustrations With LWR and CTM models. *IEEE Open Journal of Intelligent Transportation Systems*, 3:503–518, 2022.

[27] J. Zhang, S. Mao, L. Yang, W. Ma, S. Li, and Z. Gao. Physics-informed deep learning for traffic state estimation based on the traffic flow model and computational graph method. *Information Fusion*, 101:101971, 2024.

[28] B. T. Thodi, S. V. R. Ambadipudi, and S. E. Jabari. Fourier neural operator for learning solutions to macroscopic traffic flow models: Application to the forward and inverse problems. *Transportation Research Part C: Emerging Technologies*, 160:104500, 2024.

[29] B. D. Greenshields, J. R. Bibbins, W.S. Channing, and H. H. Miller. A study of traffic capacity. In *Highway Research Board proceedings*, volume 14, pages 448–477. Washington, DC, 1935.

[30] H. Greenberg. An Analysis of Traffic Flow. *Operations Research*, 7(1):79–85, 1959.

[31] U. Weidmann. Transporttechnik der Fussgänger. *Strasse und Verkehr*, 78(3):161–169, 1992.

[32] R. Poli, W. B. Langdon, and N. F. McPhee. *A Field Guide to Genetic Programming*. Lulu Enterprises UK, 2009.

[33] S. Manti and A. Lucantonio. Discovering interpretable physical models using symbolic regression and discrete exterior calculus. *Machine Learning: Science and Technology*, 5(1):015005, 2024.

[34] A. N. Hirani. *Discrete Exterior Calculus*. California Institute of Technology, 2003.

[35] E. Grinspun, M. Desbrun, K. Polthier, P. Schröder, and A. Stern. Discrete Differential Geometry: An Applied Introduction. *ACM Siggraph Course*, 7(1), 2006.

[36] R. J. LeVeque. *Finite Volume Methods for Hyperbolic Problems*, volume 31. Cambridge University Press, 2002.

[37] L. C. Edie. Car-following and Steady-State Theory for Noncongested Traffic. *Operations Research*, 9(1):66–76, 1961.

[38] P. Kachroo, S. Agarwal, A. Biswas, and A. J. Huang. Nonlocal Calculus-Based Macroscopic Traffic Model: Development, Analysis, and Validation. *IEEE Open Journal of Intelligent Transportation Systems*, 4:900–908, 2023.

[39] Y. Sun and C. Tan. On a class of new nonlocal traffic flow models with look-ahead rules. *Physica D: Nonlinear Phenomena*, 413:132663, 2020.