Socioeconomic Determinants and Mobility Patterns in COVID-19 Risk: A Spatial Network Analysis of the U.S. Counties

Holin Xue[*]
hxue49@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Priscilla Zhang[*]
zzhang3100@gatech.edu
Georgia Institute of Technology
Atanta, Georgia, USA

Abstract

Deterministic compartmental models often fail to capture the spatial connectivity and local heterogeneity arising from socioeconomic and behavioral factors, as they typically assume homogeneous mixing. To address this limitation, this study develops a spatially explicit Susceptible–Infected–Recovered (SIR) framework to analyze COVID-19 transmission dynamics across Georgia's 159 counties. Two model variants are proposed: a baseline spatial SIR model that incorporates county adjacency and case data, and an extended model that integrates socioeconomic indicators and time-varying mobility metrics obtained from open-source datasets. By systematically evaluating model performance, this study elucidates the influence of socioeconomic conditions on epidemic propagation and offers methodological guidance to enhance the design of spatially informed public health interventions.

Keywords

spatial SIR model, socioeconomic factors, mobility patterns, network epidemiology, COVID-19

1 Introduction

The spread of infectious diseases such as COVID-19 is inherently shaped by the spatial and socioeconomic characteristics of populations. Conventional deterministic compartmental models, including the classical Susceptible–Infected–Recovered (SIR) framework, typically assume homogeneous mixing, thereby overlooking heterogeneity in contact structures, mobility behaviors, and resource distributions. This simplifying assumption can yield biased predictions and obscure the role of local conditions in shaping transmission dynamics.

During the COVID-19 pandemic, Georgia experienced multiple waves of transmission that unfolded unevenly across its counties. (see Figure 1) These spatial disparities indicate that simple diffusion models based solely on geographic proximity inadequately represent the complex interactions among mobility patterns, socioeconomic conditions, and disease spread. The central problem addressed by this study is attributing socioeconomic and mobility factors to COVID-19 transmission — that is, quantifying how these structural determinants influence disease spread and identifying which factors are most predictive of transmission risk.

To address this problem, we employ a complementary two-model framework that combines mechanistic and data-driven approaches. Mechanistic SIR models are essential because they explicitly encode the causal pathway through which socioeconomic and mobility factors influence transmission rates (β_i), which in turn determine disease burden. By parameterizing transmission rates as functions of county-level covariates, these models provide interpretable coefficients that quantify the direct mechanistic impact of each factor. However, mechanistic models face limitations: linear parameterization may fail to capture complex non-linear interactions between factors, and the assumed spatial diffusion structure may not reflect the true patterns of inter-county influence.

Graph Neural Networks (GNNs) complement mechanistic models by learning flexible, non-linear relationships between socioeconomic/mobility features and disease outcomes directly from data. Unlike SIR models that impose a specific disease dynamics structure, GNNs can discover which spatial connections matter most for transmission and how features interact in ways that may not be captured by simple linear combinations. The graph structure naturally represents counties as interconnected nodes, enabling the model to learn spatial spillover effects through message passing between adjacent regions.

Both models are necessary because they address different aspects of the attribution problem. The mechanistic SIR models answer *how* socioeconomic and mobility factors influence transmission rates through a biologically-motivated framework, while the GNN answers *which* factors are most predictive and *how* they interact spatially. Together, their results provide a comprehensive view: the SIR models reveal the mechanistic pathway (SE/mobility factors → transmission rate → disease burden), while the GNN identifies the most influential factors and reveals latent spatial patterns that may not be captured by simple adjacency-based diffusion. This dual approach enables both causal interpretation (from SIR) and predictive accuracy (from GNN), ultimately contributing to a more complete understanding of how socioeconomic and mobility factors attribute to COVID-19 transmission risk.

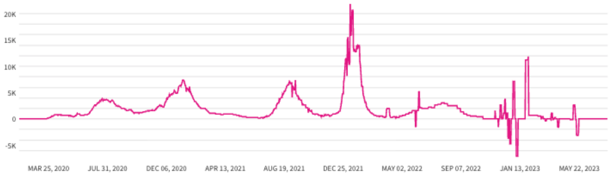

Figure 1: Daily new COVID-19 cases in Georgia, illustrating multiple waves of transmission between 2020 and 2023. Data visualization from USAFacts [5].

2 Problem Formulation

The central problem addressed by this study is attributing socioeconomic and mobility factors to COVID-19 transmission. That is,

[*] Both authors contributed equally to this research.

quantifying how these structural determinants influence disease spread and identifying which factors are most predictive of transmission risk. While numerous studies have documented correlations between socioeconomic status and disease outcomes, the mechanisms through which these factors influence disease transmission remain poorly understood, and methods for systematically attributing transmission risk to specific socioeconomic and mobility factors are limited.

Specifically, this study addresses two interconnected questions that together enable comprehensive attribution:

1. **Mechanistic attribution**: How do socioeconomic and mobility factors influence disease transmission rates (β_i) in spatially-structured epidemic models? This question seeks to quantify the causal pathway through which structural factors affect transmission.
2. **Predictive attribution**: Which socioeconomic and mobility factors are most predictive of COVID-19 burden, and how do they interact spatially? This question identifies the strongest drivers of disease risk and reveals spatial patterns of vulnerability.

To address these questions and enable comprehensive attribution, we employ a complementary two-stage analytical framework:

Stage 1 - Mechanistic Modeling (SIR): We implement spatial SIR models where the transmission rate β_i is parameterized as a function of socioeconomic and mobility covariates. This approach explicitly models the causal pathway: SE/mobility factors → transmission rate → disease burden. By calibrating the model to observed data, we learn coefficients (α_j) that quantify how each factor influences transmission rates, providing mechanistic attribution of transmission risk to specific factors.

Stage 2 - Data-Driven Modeling (GNN): Recognizing the limitations of linear parameterization in the SIR framework, we employ Graph Neural Networks to learn non-linear feature interactions and complex spatial dependencies. The GNN framework provides:

- **Predictive attribution**: Identification of which socioeconomic and mobility factors most strongly influence disease risk through variable importance analysis
- **Spatial attribution**: Learning which spatial connections matter most for disease spread, revealing how neighboring counties influence each other
- **Structural insights**: Clustering of counties based on learned embeddings reveals latent groups with similar risk profiles, helping explain how SE and mobility factors jointly influence risk

3 Literature Review

To inform our study, we conducted a literature review on topics including socioeconomic disparities in COVID-19 outcomes and spatial interrelation models for human mobility. This review provided insights into how social and spatial factors jointly influence disease transmission and informed the development of our analytical framework.

3.1 Socioeconomic Disparities in COVID-19

The COVID-19 pandemic has amplified pre-existing health inequities, particularly those rooted in socioeconomic status (SES). SES - encompassing income, education, occupation, and housing quality—has long been a determinant of health outcomes, influencing exposure risk, healthcare access, and disease severity. During the pandemic, communities with lower SES consistently experienced higher rates of infection, hospitalization, and mortality, revealing how structural disadvantage can shape vulnerability in times of crisis.

Khanijahani et al. [3] synthesized evidence from numerous studies examining how socioeconomic factors such as education, income, housing, and language barriers influenced COVID-19 outcomes. Their review concluded that populations with lower SES were disproportionately affected across all metrics of disease burden. Similarly, Wachtler et al. [6], in a review of 138 studies from the United States, United Kingdom, and Europe, found consistent evidence that low-income and otherwise disadvantaged populations faced elevated risks of SARS-CoV-2 infection and more severe outcomes. Together, these findings underscore the strong link between socioeconomic disadvantage and COVID-19 incidence, yet most existing studies focus on descriptive associations rather than predictive modeling. To bridge this gap, we propose a network-based predictive framework to assess whether integrating socioeconomic variables improves model accuracy and interpretability.

Beyond establishing correlation, recent work has begun to explore mechanisms underlying these disparities. Chang et al. [2] developed an SEIR-based mobility network model using anonymized cellphone data from ten major U.S. metropolitan areas to simulate COVID-19 transmission dynamics. Their simulations accurately reproduced the higher infection rates observed among low-income populations without invoking biological or health-based differences. Instead, they attributed these disparities to structural mobility constraints—such as essential employment, reliance on public transportation, and greater exposure in crowded venues. These findings highlight that socioeconomic disparities in COVID-19 outcomes stem not from individual behavior alone, but from systemic social and structural inequities that constrain the ability to mitigate exposure risk.

3.2 Spatial Interrelation Models for Human Mobility

Early work by Balcan and Vespignani [1] established one of the foundational frameworks for analyzing contagion processes mediated by recurrent mobility. By coupling a metapopulation structure with the SIR model, they demonstrated that bidirectional commuting patterns create a critical invasion threshold separating localized from system wide outbreaks. Their results revealed that both the diffusion rate and return rate of travelers jointly determine whether a contagion remains spatially confined or becomes widespread. Although this framework provided a theoretical basis for understanding mobility-driven epidemics, it assumed homogeneous travel behavior and did not explicitly incorporate socioeconomic or infrastructural heterogeneity.

Subsequent empirical studies began testing the generalizability of such theoretical mobility models in real-world settings. Wesolowski et al. [7] assessed the performance of gravity and radiation models in representing regional mobility patterns in Sub-Saharan Africa using anonymized mobile phone data from approximately 15 million Kenyan subscribers. Their findings revealed systematic biases: the gravity model consistently overestimated travel flows, while the radiation model underestimated them, particularly in rural areas. These discrepancies highlighted the inadequacy of universal mobility formulations in contexts with strong infrastructural constraints and informal movement patterns. The study underscored the need for region-specific models that account for socioeconomic and geographic diversity when linking human mobility to disease transmission.

More recent work has incorporated detailed spatial and behavioral data to refine these earlier models. Tokey [4] analyzed the association between mobility and COVID-19 infection rates across U.S. counties using GPS-based mobility data from March to August 2020. Employing spatial regression approaches, including Ordinary Least Squares (OLS), Spatial Error Models (SEM), and Geographically Weighted Regression (GWR), the study showed that infection rates were inversely correlated with mobility indicators such as miles traveled and out-of-county trips. Spatial models captured geographic heterogeneity in transmission more effectively than non-spatial alternatives, though the use of aggregate data limited insights into individual behavioral mechanisms. This study advanced the integration of spatial econometrics into epidemic modeling, demonstrating the empirical importance of spatial non-stationarity in understanding COVID-19 diffusion.

Building upon this body of work, the present study applies a spatially SIR framework to Georgia's 159 counties, integrating time-varying mobility and socioeconomic indicators to examine how local heterogeneity drives COVID-19 transmission and to improve the predictive and explanatory power of spatial epidemic models.

4 Data

We compiled a county-level dataset encompassing 3,222 U.S. counties to examine how socioeconomic context and mobility patterns shaped COVID-19 risk. Weekly COVID-19 outcomes—including incidence, hospital admissions, and inpatient bed utilization—were obtained from the CDC COVID-19 Community Levels dataset. County adjacency data defined the spatial network structure. Socioeconomic indicators, sourced from the U.S. Census Bureau, capture income, education, employment, housing, and demographic composition (the full list of 18 socioeconomic variables inluded in this study is presented in Table 1 ih Appendix A). Mobility data from the Google COVID-19 Community Mobility Reports quantify relative changes in visits across six categories of places. COVID-19 burden is summarized as the proportion of weeks each county was classified at high risk, providing a normalized measure of sustained disease intensity.

5 Methods

5.1 Spatial SIR Model Framework

We employ a spatially-structured SIR model to simulate COVID-19 transmission across U.S. counties. Each county i is represented as a node in a weighted, undirected network $G = (V, E)$, where edges encode spatial adjacency between counties. The model is defined as:

$$\frac{dS_i}{dt} = -\beta_i S_i \sum_{j \in \mathcal{N}(i)} w_{ij} \frac{I_j}{N_j} + m \sum_{j \in \mathcal{N}(i)} w_{ij} \left(\frac{S_j}{N_j} - \frac{S_i}{N_i} \right)$$

$$\frac{dI_i}{dt} = \beta_i S_i \sum_{j \in \mathcal{N}(i)} w_{ij} \frac{I_j}{N_j} - \gamma I_i + m \sum_{j \in \mathcal{N}(i)} w_{ij} \left(\frac{I_j}{N_j} - \frac{I_i}{N_i} \right)$$

$$\frac{dR_i}{dt} = \gamma I_i + m \sum_{j \in \mathcal{N}(i)} w_{ij} \left(\frac{R_j}{N_j} - \frac{R_i}{N_i} \right)$$

where S_i, I_i, R_i denote the susceptible, infected, and recovered populations in county i, N_i is the total population of county i, $\mathcal{N}(i)$ represents the set of neighboring counties, w_{ij} represents spatial connectivity (normalized by node degree for diffusion terms), β_i is the county-specific transmission rate, γ is the recovery rate (assumed constant across counties), and m is the spatial diffusion parameter controlling cross-county movement.

Initial conditions: For each county, initial conditions are set based on observed case data: I_0 = observed cases, $S_0 = N - I_0$, and $R_0 = 0$. The model is simulated for $T = 60$ time steps to reach steady-state conditions.

5.2 SIR Model Variants

Two model variants are constructed to assess the influence of socioeconomic and mobility heterogeneity:

5.2.1 Baseline Spatial SIR Model. The Baseline Spatial SIR model incorporates only county adjacency relationships and reported case data to capture baseline spatial diffusion through direct geographic connectivity. The transmission rate is constant across all counties: $\beta_i = \beta$ for all i. This model serves as a baseline to evaluate whether incorporating socioeconomic heterogeneity improves predictive performance.

Parameters to calibrate:

- β: Constant transmission rate (bounded: $[0.001, 0.1]$)
- γ: Recovery rate (bounded: $[1/21, 1/3]$ days^{-1}, corresponding to infectious periods of 3-21 days)
- m: Spatial diffusion parameter (bounded: $[0, 0.2]$)

5.2.2 Extended Socio-mobility Model. The Extended Socio-mobility model augments the baseline by allowing β_i to vary across counties as a function of socioeconomic and mobility covariates. The transmission rate is parameterized as:

$$\beta_i = \exp\left(\alpha_0 + \sum_{j=1}^{p} \alpha_j \cdot X_{ij} \right)$$

where X_{ij} represents the j-th standardized socioeconomic feature for county i, α_0 is the intercept, and α_j are feature-specific coefficients. The exponential transformation ensures $\beta_i > 0$ for all counties, and the resulting values are clipped to $[0.001, 0.1]$ to maintain realistic transmission rates.

Note on time-varying parameters: While the proposal originally envisioned time-varying $\beta_i(t)$, the current implementation uses static β_i values due to data limitations (using latest snapshot per county). The framework is designed to accommodate time-varying parameters when temporal data becomes available.

Parameters to calibrate:

- $\alpha_0, \alpha_1, \ldots, \alpha_p$: Feature weights and intercept ($p = 18$ socioeconomic features)
- γ: Recovery rate (same bounds as baseline)
- m: Spatial diffusion parameter (same bounds as baseline)

Calibration process: Both models are calibrated using optimization (L-BFGS-B) to minimize Mean Absolute Error (MAE) between predicted and observed disease burden (cases per 100,000 population). The optimization process simultaneously adjusts all parameters to find values that best reproduce observed disease patterns.

5.3 Graph Neural Network Model

The Extended Socio-mobility SIR model's linear parameterization revealed several limitations that motivate a more flexible approach. Additionally, the simple spatial diffusion mechanism treats all neighbor connections equally, unable to learn which spatial relationships matter most for disease spread.

To address these limitations and provide complementary attribution capabilities, we employ a Graph Neural Network (GNN) that represents counties as interconnected nodes in a spatial network. Each county is characterized by its socioeconomic and mobility features, while edges connect adjacent counties, creating a graph structure that captures the spatial relationships between regions. The graph structure enables the model to capture spatial dependencies through message passing, where information flows between neighboring counties, allowing the model to learn which spatial connections are most predictive of disease burden.

The GNN serves two complementary objectives for attribution. First, it predicts disease burden from socioeconomic and mobility features while accounting for spatial spillover effects: how neighboring counties influence each other. This addresses the attribution problem by quantifying how socioeconomics (SE) and mobility factors jointly shape COVID-19 risk, with spatial dependencies learned from the data rather than assumed. Second, the model provides interpretable attribution insights through two mechanisms: (1) variable importance analysis identifies which socioeconomic and mobility factors most strongly influence predictions, enabling attribution of disease risk to specific factors, and (2) embedding-based clustering groups counties with similar risk profiles, revealing latent structural patterns (e.g., clusters characterized by low income but high disease burden) that help explain the nature of how SE and mobility factors influence risk.

6 Experiment Design

To address the research question of how socioeconomic characteristics and mobility patterns influence COVID-19 risk, we employ a comprehensive experimental framework with explicit baselines, ablation studies, and clear evaluation metrics.

6.1 Baseline Models

Three baseline models are established to provide comparison points:

1. **Baseline Spatial SIR**: Constant transmission rate β across all counties, incorporating only spatial adjacency and case data. This serves as the simplest mechanistic model.
2. **Extended Socio-mobility SIR**: County-specific transmission rates β_i parameterized as a linear combination of socioeconomic features. This tests whether simple linear parameterization improves upon the baseline.
3. **Ridge Regression**: A non-spatial linear regression baseline that treats counties independently, providing a comparison for the importance of spatial structure.

6.2 Ablation Studies

To understand the contribution of different components, we conduct the following ablation studies:

1. **Spatial vs. Non-spatial**: Compare models with and without spatial structure (adjacency network) to quantify the importance of spatial dependencies.
2. **With vs. Without SE features**: Compare Baseline SIR (no SE features) vs. Extended SIR (with SE features) to quantify the contribution of socioeconomic heterogeneity.
3. **Linear vs. Non-linear**: Compare Extended SIR (linear parameterization) vs. GNN (non-linear) to assess whether non-linear feature interactions improve performance.
4. **Feature importance analysis**: Within the Extended SIR and GNN models, analyze which socioeconomic features receive the largest weights/importance, providing insights into which factors most strongly influence disease risk.

6.3 Evaluation Framework

The following metrics are used to evaluate the models.

- **Mean Absolute Error (MAE)**: Primary metric for prediction accuracy
- **Root Mean Squared Error (RMSE)**: Captures larger errors more heavily
- **Coefficient of Determination (R^2)**: Measures proportion of variance explained

6.4 Expected Outcomes

Based on the experimental design, we expect to find:

- SE factors that correlate with disease burden (direct relationships)
- SE factors that influence transmission rates (mechanistic pathway)
- Agreement/disagreement between correlations and transmission mechanisms
- Improved predictive performance with GNN compared to linear SIR models

7 Results

7.1 SIR Results

7.1.1 Baseline Spatial SIR Model. The Baseline Spatial SIR model incorporates only county adjacency relationships and reported case data, with a constant transmission rate β across all counties. Parameter calibration via optimization yielded the following estimates:

- Transmission rate (β): 0.001 (constant across all counties)
- Recovery rate (γ): 0.333, corresponding to an infectious period of approximately 3.0 days
- Spatial diffusion parameter (m): 0.019981

The model's predictive performance, evaluated on cases per 100,000 population, is summarized as follows:

- Mean Absolute Error (MAE): 1.26 cases per 100,000
- Root Mean Squared Error (RMSE): 7.65 cases per 100,000
- Coefficient of Determination (R^2): -0.028

The negative R^2 value indicates that the baseline model performs worse than a simple mean predictor, suggesting that the homogeneous transmission rate assumption fails to capture the heterogeneity in disease burden across counties. The low MAE (1.26) relative to RMSE (7.65) suggests that while most predictions are reasonably close to observed values, the model produces occasional large errors.

7.1.2 Extended Socio-mobility Model. The Extended Socio-mobility model augments the baseline by allowing the transmission rate β_i to vary across counties as a function of socioeconomic covariates. The model parameterizes β_i as:

$$\beta_i = \exp\left(\alpha_0 + \sum_{j=1}^{18} \alpha_j \cdot \text{feature}_{ij}\right)$$

where the 18 features include socioeconomic indicators such as income, education, poverty, housing conditions, and demographic composition. Note the model incorporates only socioeconomic features for now.

Parameter calibration yielded county-specific transmission rates with the following characteristics:

- Transmission rate range: $\beta_i \in [0.001, 0.100]$ (clipped to ensure realistic values)
- Mean transmission rate: $\bar{\beta} = 0.1$
- Recovery rate (γ): 0.142857, corresponding to an infectious period of approximately 7.0 days
- Spatial diffusion parameter (m): 0.02

The Extended model's predictive performance is as follows:

- Mean Absolute Error (MAE): 13.53 cases per 100,000
- Root Mean Squared Error (RMSE): 17.88 cases per 100,000
- Coefficient of Determination (R^2): -4.609

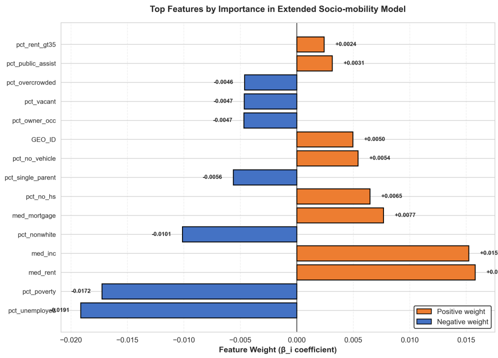

Figure 2: Selected Feature Importance

7.1.3 Model Comparison and Feature Analysis. Figure 2 visualizes these feature weights, showing both the magnitude and direction of each socioeconomic factor's influence on transmission rates. The predominance of negative weights for traditionally disadvantaged indicators (unemployment, poverty) is counterintuitive and may reflect the limitations of the linear parameterization approach.

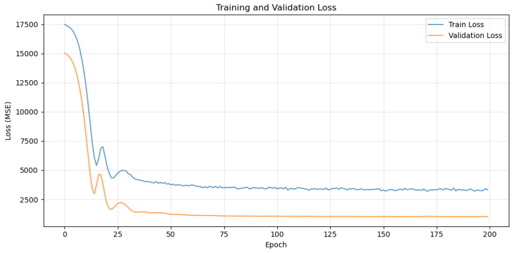

Figure 3: Training and Validation Loss of GNN

7.2 GNN Results

7.2.1 GNN Model Performance. The Graph Neural Network (GNN) was trained using socioeconomic and mobility features as node attributes and county adjacency as the graph structure. In contrast to the SIR models, the GNN directly learned spatial and feature-level dependencies through message passing rather than imposing them analytically.

The model exhibited stable optimization dynamics as shown in Figure 3: both training and validation losses decreased steeply during early epochs and gradually plateaued, indicating good convergence without signs of overfitting.

The GNN model's predictive performance is shown below:

- Mean Absolute Error (MAE): 27.48 cases per 100,000
- Root Mean Squared Error (RMSE): 39.59 cases per 100,000
- Coefficient of Determination (R^2): 0.441

Although the MAE and RMSE of the GNN are numerically larger than those of the SIR baselines due to differing scale of predicted targets, the GNN is the only model that achieves positive explanatory power, with $R^2 = 0.441$. Predictions closely align with the 45° reference line for most counties (Figure: Predictions vs Actual), while the residual plot shows no systematic bias across the predicted range. These results confirm that the GNN captures spatial spillover effects and heterogeneous socioeconomic influences that the analytic SIR models cannot represent.

7.2.2 Feature Importance. Permutation-based feature importance reveals the strongest predictors of COVID-19 burden. The highest scoring variables include:

- Median household income
- Residential mobility (percent change from baseline)
- Percent of single-parent households
- Median mortgage cost
- Overcrowding rate
- Percent without high school diploma
- Workplace mobility (percent change from baseline)

The full list of feature importance is showns in Figure 4 in Appendix A.

The prominence of income, education, housing costs, and household structure indicates that structural socioeconomic vulnerability remains a dominant driver of disease risk, even after controlling for mobility. In addition, both residential and workplace mobility appear among the most influential predictors, confirming that behavioral movement patterns mediate the link between socioeconomic disadvantage and transmission intensity.

8 Discussion

This study set out to address the problem of attributing socioeconomic and mobility factors to COVID-19 transmission by quantifying how these structural determinants influence disease spread across U.S. counties. Through a complementary framework combining mechanistic and data-driven approaches, we provide both causal attribution (how factors influence transmission rates) and predictive attribution (which factors are most influential). The findings demonstrate that while structural socioeconomic disadvantage is strongly associated with elevated COVID-19 burden, the ability of modeling frameworks to make use of this information critically depends on their representational flexibility.

The comparison between the two SIR model variants highlights the limitations of linear mechanistic parameterization. The Baseline Spatial SIR model, which assumes a homogeneous transmission rate and relies solely on geographic adjacency, failed to reproduce observed spatial heterogeneity in disease burden ($R^2 = -0.028$). The Extended Socio-mobility SIR model attempted to incorporate socioeconomic heterogeneity by linking transmission rates to county-level covariates. However, despite the theoretical motivation for this mechanistic pathway, model performance deteriorated sharply ($R^2 = -4.609$). In addition, the predominance of negative coefficients for variables traditionally associated with higher risk—such as poverty, unemployment, and lack of education—suggests that simple linear parameterization is structurally misaligned with the true dependence of transmission dynamics on socioeconomic conditions. Taken together, these results indicate that although socioeconomic inequity is an important determinant of disease risk, its effects are highly non-linear, interact with other drivers such as mobility, and cannot be captured through an additive exponential formulation.

By contrast, the Graph Neural Network provided substantially improved explanatory power ($R^2 = 0.441$), confirming that non-linear spatial and socioeconomic dependencies play a central role in shaping COVID-19 burden. The convergence behavior of the training and validation losses and the absence of systematic bias in the residual plot demonstrate that the GNN was able to learn generalizable patterns rather than memorizing local noise. Feature importance estimates provide further evidence that structural inequity remains a dominant driver of disease outcomes: median household income, overcrowding, percent without a high school diploma, mortgage burden, and single-parent households emerged among the strongest predictors. At the same time, mobility-derived indicators—especially changes in residential and workplace activity—also ranked highly, suggesting that mobility mediates the link between socioeconomic disadvantage and infection exposure. These insights align with recent epidemiological evidence that essential work requirements, limited remote-work capacity, and reliance on shared transportation disproportionately increase exposure risk in disadvantaged counties.

Importantly, the improvement in predictive performance achieved by the GNN does not imply that mechanistic epidemic modeling is obsolete. Rather, our results point toward an opportunity for hybrid modeling. The SIR framework provides interpretability and a causal lens on transmission mechanisms but struggles with high-dimensional heterogeneity; deep learning models achieve high predictive power but do not explicitly encode disease dynamics. A promising direction for future work is the integration of GNNs into mechanistic epidemic models. For example, using message-passing networks to learn spatiotemporal contact structures or to parameterize transmission rates in a flexible and data-driven way.

Overall, this study provides empirical evidence that enables attribution of COVID-19 transmission risk to specific socioeconomic and mobility factors. Through the complementary use of mechanistic and data-driven models, we demonstrate that the intersection of socioeconomic inequality and mobility drives the spatial diffusion of COVID-19, and that flexible spatial learning frameworks are required to capture these mechanisms. The dual-model approach enables comprehensive attribution: mechanistic models reveal how factors influence transmission rates through causal pathways, while the GNN identifies which factors are most predictive and how they interact spatially. Counties experiencing greater structural disadvantage face disproportionate disease burden, not solely because of biological vulnerability but because socioeconomic and mobility constraints intensify exposure and reduce ability to mitigate risk.

The attribution framework developed here, combining mechanistic understanding with predictive identification, provides actionable insights for public health policy. By identifying which factors most strongly influence transmission (income, education, housing, mobility) and understanding how they mechanistically affect disease spread, this work enables targeted interventions that address structural determinants rather than only biomedical factors. These findings reinforce the need for public health policies that target not only biomedical interventions but also structural determinants, such as safe working conditions, housing security, and mobility access—to mitigate inequitable pandemic outcomes.

References

[1] Duygu Balcan and Alessandro Vespignani. 2011. Phase transitions in contagion processes mediated by recurrent mobility patterns. *Nature Physics* 7, 7 (2011), 581–586. doi:10.1038/nphys1944

[2] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. 2021. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature* 589, 7840 (2021), 82–87.

[3] Ahmad Khanijahani, Shabnam Iezadi, Kamal Gholipour, Saber Azami-Aghdash, and Deniz Naghibi. 2021. A systematic review of racial/ethnic and socioeconomic disparities in COVID-19. *International Journal for Equity in Health* 20, 1 (2021), 248.

[4] Ahmad Ilderim Tokey. 2021. Spatial association of mobility and COVID-19 infection rate in the USA: A county-level study using mobile phone location data. *Journal of Transport & Health* 22 (2021), 101135. doi:10.1016/j.jth.2021.101135

[5] USAFacts. 2025. *Georgia coronavirus cases and deaths.* USAFacts. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/state/georgia/ Accessed: 2025-10-08.

[6] Benjamin Wachtler, Niels Michalski, Enno Nowossadeck, Michaela Diercke, Morten Wahrendorf, Claudia Santos-Hövener, Thomas Lampert, and Jens Hoebel. 2020. Socioeconomic inequalities and COVID-19 –A review of the current international literature. S7 (2020), 3–17.

[7] Amy Wesolowski, Wendy Prudhomme O'Meara, Nathan Eagle, Andrew J. Tatem, and Caroline O. Buckee. 2015. Evaluating Spatial Interaction Models for Regional Mobility in Sub-Saharan Africa. *PLOS Computational Biology* 11, 7 (2015), e1004267. doi:10.1371/journal.pcbi.1004267

A Appendix - Socioeconomic Variables

The following table presents a complete list of 18 socioeconomic variables included in this study.

Table 1: Socioeconomic Variables and Corresponding Census Table IDs

Variable Name	Table ID
Percent population aged 25 and above without a High School diploma	S1501
Median household income in US dollars	S1903
Income disparity (Gini Index)	B19083
Median home value in US dollars	DP04
Median gross rent in US dollars	DP04
Median monthly mortgage in US dollars	DP04
Percent of owner-occupied housing units	DP04
Percent of civilian labor force population aged 16 years and older who are unemployed	S2301
Percent of families below federal poverty level	S1702
Percent of single-parent households with children less than 18 years of age	DP02
Percent of households without a motor vehicle	DP04
Percent of households with more than 1 person per room	DP04
Percent non-white	DP05
Percent of vacant housing units	DP04
Percent of households where rent is greater than 35% of household income	DP04
Percent of households receiving public assistance income	B19057
Percent of individuals without health insurance	S2701
Percent of households without internet access	S2801

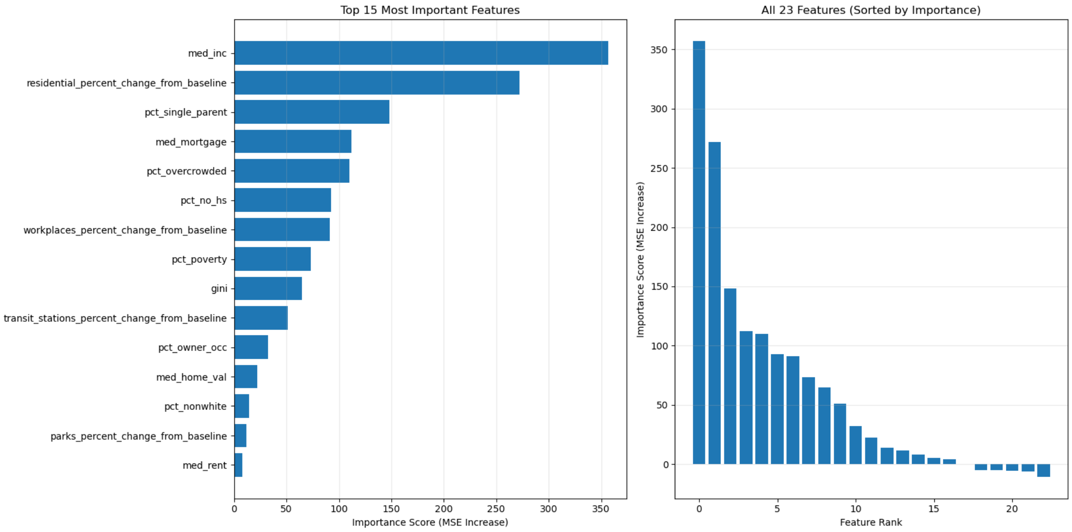

Figure 4: GNN Feature Importance