Appendix for Back2Future: Leveraging Backfill Dynamics for Improving Real-time Predictions in Future

Code for B2F and the *CoVDS* dataset is publicly available[2]. Please refer to the README in the code folder for more details to reproduce the results.

A Additional Related work

Epidemic forecasting. Broadly two classes of approaches have been devised: traditional mechanistic epidemiological models Shaman & Karspeck (2012); Zhang et al. (2017), and the fairly newer statistical approaches Brooks et al. (2018); Adhikari et al. (2019); Osthus et al. (2019b), which have become among the top performing ones for multiple forecasting tasks Reich et al. (2019).

Statistical models have been helpful in using digital indicators such as search queries Ginsberg et al. (2009); Yang et al. (2015) and social media Culotta (2010); Lampos et al. (2010), that can give more lead time than traditional surveillance methods. Recently, deep learning models have seen much work. They can use heterogeneous and multimodal data and extract richer representations, including for modeling spatio-temporal dynamics Adhikari et al. (2019); Deng et al. (2020) and transfer learning Panagopoulos et al. (2021); Rodríguez et al. (2021a). Our work can be thought of refining any model (mechanistic/statistical) in this space.

Revisions and backfill. The topic of revisions has not received as much attention, with few exceptions. In epidemic forecasting, a few papers have mentioned about the 'backfill problem' and its effects on performance (Chakraborty et al., 2018; Rodríguez et al., 2021b; Altieri et al., 2021; Rangarajan et al., 2019) and evaluation (Reich et al., 2019); Some works proposed to address the problem via simple models like linear regression (Chakraborty et al., 2014) or 'backcasting' (Brooks et al., 2018) the observed targets. However they focus only on revisions in the *target*, and study only in context of influenza forecasting, which is substantially less noisy and more regular than forecasting for the novel COVID-19 pandemic. Reich et al. (2019) proposed a framework to study how backfill affects the evaluation of multiple models, but it is limited to label backfill and flu forecasting. Other works use data assimilation and sensor fusion by leveraging revision free digital signals to refine noisy features for nowcasting (Hawryluk et al., 2021; Farrow, 2016; Lampos et al., 2015; Nunes et al., 2013; Osthus et al., 2019a). However, we observed significant backfill in digital signals as well for COVID-19. Moreover, our model doesn't require revision-free data sources. Some works model revision of event-based features as count variables which can't be applicable to many important features like mobility, exposure (Stoner & Economou, 2020; Lawless, 1994). Clements & Galvão (2019) surveys several domain-specific (Carriero et al., 2015) or essentially linear techniques in economics for data revision/correction behavior of the source of several macroeconomic indicators (Croushore, 2011). In contrast, we study the more challenging problem of multi-variate backfill for both features and targets and show how to leverage our insights for more general neural framework to *improve* both model predictions and evaluation.

B Details about GDP prediction task

B.1 Dataset

We curated the dataset collected from the Real-Time Data Research Center, Federal Reserve Bank Philadelphia which can be accessed publicly [3]. We extracted the following important macroeconomic features denoted by tickers in parenthesis: Real GDP (ROUTPUT), Real Personal Consumption Expenditures Goods and Services(RCON, RCONG, RCONS), Real Gross Private Domestic Investment (RINVBF, RINVRESID), Real net Exports (RNX), Total Government Consumption (RG, RGF, RGSL), Household Consumption Expenditure (RCONHH), Final Consumption Expenditure (NCON), Nominal GDP (NOUTPUT), Nominal Personal Consumption Expenditures (NCON), Wage and Salary Disbursement (WSD), Other labor Income (OLI), Rental Income (RENTI), Dividends (DIV), Personal Income (PINTI), Transfer Payments (TRANR), Personal Saving Rate (RATESAV),

[2] https://github.com/AdityaLab/Back2Future

[3] Link to dataset: https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/real-time-data-set-full-time-series-history

Disposable Personal Income (NDPI), Unemployment Rate (RUC), Civilian Labor Force (LFC), Consumer Price Index (CPI).

B.2 SETUP

For each of these 25 features, quarterly data is available since 1965 to third quarter of 2021 and the data is revised quarterly towards more accurate values. Our goal is to predict the Real GDP (ROUTPUT) k quarters ahead using real-time data (data available till current quarter including past data revised till current quarter) where $k \in \{1, 2\}$. We tune the hyperparameters for model for using data from years 1995-2000 and test on the unseen period of 2000-2021. Due to lack of standard baselines, we chose as candidate models 1. Vector Autoregression (VAR) model, a standard model in macroeconmics literature (Robertson & Tallman, 1999; Rünstler & Sédillot, 2003; Baffigi et al., 2004), 2. 4-layer feed forward network (FFN) on current quarter's features similar to (Tkacz & Hu, 1999) and a Recurrent Neural Network (GRU) on past GDP values.

C NATURE OF BACKFILL DYNAMICS (MORE DETAILS)

C.1 CONSIDERING DELAY IN REPORTING

Let current week be t. Due to delay in reporting, the first observed value for some signals could be delayed by 1 to 3 weeks. For instance, $\text{BSEQ}(i,t)$ may not have the first δ_i values due to δ_i weeks delay in reporting. In such cases, for analysis of BSEQ in Section 2, we take $d_{i,t}^{t+\delta_i}$ as the first value of $\text{BSEQ}(i,t) = \langle d_{i,t}^{t+\delta_i}, d_{i,t}^{t+\delta_i+1}, \cdots, d_{i,t}^{t_f} \rangle$. Subsequently, $\text{BERR}(r,i,t)$ is only defined for $r > \delta_i$ as $\text{BERR}(r,i,t) = \frac{|d_{i,t}^{(t+r)} - d_{i,t}^{t_f}|}{|d_{i,t}^{t_f}|}$.

Real-time forecasting As mentioned in the main paper, we handle delays in reporting real-time data by approximating from the previous week. During real-time forecasting, we cannot wait for δ_i weeks to get the first value of a signal. Therefore, we replace $d_{i,t}^{(t)}$ with the most revised value of last week for which the signal is available. Let $t' < t$ is the last week before t for which we have $d_{i,t'}^{(t)}$. Then, we use $d_{i,t'}^{(t)}$ in place of unavailable $d_{i,t}^{(t)}$. Note that such cases of missing initial value during real-time forecasting is very uncommon.

C.2 DESCRIPTION OF CANONICAL BACKFILL BEHAVIOURS

We saw in Observation 3 that clustering BSEQs resulted in 5 canonical behaviours. The behaviors of these clusters as shown in Figure 2 can be described as:

1. Early Decrease: BSEQ values stabilizes quickly (within a week) to a lower stable value.
2. Early Increase: BSEQ values increase in 2 to 5 weeks and stabilize.
3. Steady/Spike: BSEQ values either remain constant (no significant revision) or due to reporting errors there may be anomalous values in between. (E.g., for a particular week in the middle of BSEQ the signals are revised to 0 due to reporting error).
4. Late Increase: BSEQ values increase and stabilize very late during revision.
5. Mid Decrease: BSEQ values change and stabilize after 8 to 10 weeks of revision.

Table 3: Pearson Correlation Coefficient (PCC) between BERR and REVDIFFMAE using stable labels over real-time labels

Model	**PCC**
ENSEMBLE	-0.327
GT-DC	-0.322
YYG	0.110
UMASS-MB	-0.291
CMU-TS	-0.468

C.3 CORRELATION BETWEEN BERR AND MODEL PERFORMANCE

In Observation 4, we found that models differ in how their performance is impacted by BERR on the label and found that relationship between BERR and difference in MAE (REVDIFFMAE) was varied across models with some models (YYG) even showing positive correlation between BERR and REVDIFFMAE. To further study this relationship between BERR and reduction in MAE on using *stable* labels for evaluation over *real-time* labels by computing the Pearson correlation coefficient (PCC) between BERR and REVDIFFMAE for in Table 3. As seen from Figure 3, we observe that for YYG, PCC is positive, indicating that YYG's scores are actually due to larger BERR on labels. We also see significant differences in PCC for YYG and CMU-TS in over other models.

D HYPERPARAMETERS

In this section, we describe in detail hyperparameters related to data preprocessing and B2F architecture.

D.1 DATA PRE-PROCESSING

Aggregating Google mobility features For each state-level mobility feature (`RetailRec`, `Grocery`, `Parks`, `Transit`, `WorkSpace` and `Resident`), we aggregate the percentage increase in mobility reported for each county as well as for entire state. This captures the combined revision dynamics at community and state-level.

Missing real-time data Sometimes there is a delay in receiving signals for the current week. In this case, we use values from the most revised version of last observation week for real-time forecasting.

Missing revisions Once we start observing a signal $d_t^{(t)}$ from week t till t_f, there are weeks t' in between where the revised value of this signals is not available or value received is zero. This gives rise to *Spike* behaviour (Figure 2). Before training, we replace this value with the previous value of BSEQ.

Termination of revisions We observed that for some digital signals, revisions stop a few months in the future. This could be due to the termination of data correction for that signal for older observation weeks. In such cases, we assume that signals are stabilized and use the last available revised values to fill the following values of BSEQ.

Scaling signal values Since each signal that is received has a very different range of values, we rescale each signal with a mean 0. and standard deviation 1.0. Note that since B2F is trained separately for each week, we normalize the data for each week before training.

D.2 ARCHITECTURE

BSEQENC The dimension size of all latent encodings $h_{i,t'}^{(t'_r)}$ and $v_{i,t'}^{(t'_r)}$ is set to 50. Each $GConv$ is a single layer of graph convolutional neural network with weight matrix of size $\mathrm{R}^{50\times 50}$.

MODELPREDENC For GRU_{ME} We use a GRU of a single hidden layer with output size 50.

REFINER the feed-forward network $FFN_{RF}$is a 2 layer network with hidden layers of size 60 and 30 followed by final layer outputting 1-dimensional scalar.

Training hyperparameters We used a learning rate of 10^{-3} for pre-training and 5×10^{-4} for fine-tuning. The pretraining task usually around takes 2000 epoch to train with the first 1000 epochs using teacher forcing and each of the rest of the epoch using teacher forcing with a probability of 0.5. Fine-tuning takes between 500 to 1000 epoch depending on the model, region, and week of the forecast.

As mentioned in Section 4, we used data from June 2020 to Dec 2020 for model design using Observations from Section 2. The hyperparameter tuning was done using data from June 2020 to Aug 2020 and evaluated for time period of Jan 2021 to July 2021. Overall, we found that most hyperparameters are not sensitive. The most sensitive ones mentioned in the main paper are $c \in \{2, 3, 4, 5\}$ that controls sparsity of graph G_t and $l = 5$ that controls how many steps we auto-regress using BSEQENC to derive latent encodings for BSEQ during inference.

E ADDITIONAL RESULTS

E.1 COVID-19 FORECASTING

We show the average % improvements of all baselines and B2F in Table 4 including for $k = 1, 3$ week ahead forecasts (We show for $k = 2, 4$ in main paper Table 2 as well). We show the results for the both June 2020 to Dec 2020 data, which was observed during model design and Jan 2021 to June 2021 which was unseen. B2F show similar performance for both time periods. B2F clearly outperforms all baselines and provides similar improvements for COVID-19 Forecast Hub models for $k = 1$ week ahead forecasts as described in Section 4 for other values of k.

Table 4: % improvement in MAE and MAPE scores averaged over all regions from June 2020 to Dec 2020

		k=1		k=2		k=3		k=4	
Cand. Model	Refining Model	**MAE**	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**
ENSEMBLE	FFN	-0.21	-0.45	-0.35	-0.12	0.48	-0.29	0.81	0.36
	B2F-MB	-1.67	-0.61	-2.23	-1.57	-1.51	-3.78	-1.13	-2.36
	B2F-NOGRAPH	0.15	0.19	-1.45	-2.73	-1.74	-3.21	-5.2	-4.78
	B2F-BSEQ	2.1	0.29	1.42	0.37	0.37	0.22	0.72	0.28
	B2F	**6.22**	**6.13**	**5.18**	**5.47**	**3.64**	**3.9**	**3.61**	**6.3**
GT-DC	FFN	-1.92	-1.45	-2.42	-1.51	-2.04	-6.07	-1.86	-0.49
	B2F-MB	-3.9	-4.44	-3.02	-3.41	-2.62	-3.95	-2.81	-3.26
	B2F-NOGRAPH	0.51	0.46	2.24	3.51	2.31	2.42	1.6	0.57
	B2F-BSEQ	2.92	2.5	2.13	3.84	3.6	3.55	1.42	2.27
	B2F	**10.44**	**12.37**	**12.68**	**12.59**	**11.42**	**10.67**	**9.64**	**8.97**
YYG	FFN	-3.45	-1.53	-2.08	-1.34	-4.62	-3.89	-1.37	-3.06
	B2F-MB	-1.47	-0.92	-3.84	-6.99	-5.57	-4.16	-9.29	-5.81
	B2F-NOGRAPH	-1.39	-0.59	-1.25	-0.7	-2.57	-3.72	-6.65	-5.45
	B2F-BSEQ	-1.98	-1.92	-1.78	-2.26	-1.99	-1.53	-0.61	-0.43
	B2F	**10.64**	**7.74**	**8.84**	**5.98**	**7.04**	**7.64**	**6.8**	**5.27**
UMASS-MB	FFN	-2.37	-2.03	-3.25	-5.74	-2.16	-3.17	-1.44	-4.84
	B2F-MB	-3.52	-4.61	-8.2	-7.54	-5.19	-7.82	-5.99	-7.43
	B2F-NOGRAPH	-1.01	-0.92	-2.16	-1.88	-2.13	-0.67	-2.29	-2.31
	B2F-BSEQ	0.92	0.78	1.58	0.86	-1.26	0.45	0.06	0.03
	B2F	**4.44**	**5.31**	**5.21**	**4.92**	**3.25**	**4.74**	**3.94**	**3.49**
CMU-TS	FFN	-2.22	-4.17	-5.24	-4.93	-2.87	-1.95	-3.19	-6.7
	B2F-MB	-6.59	-4.11	-8.17	-8.21	-3.32	-6.3	-3.75	-9.1
	B2F-NOGRAPH	0.46	0.71	-0.67	-0.57	-0.73	-0.19	-0.38	-0.12
	B2F-BSEQ	1.76	2.34	1.46	1.05	2.74	2.47	2.11	2.84
	B2F	**7.54**	**8.22**	**8.75**	**10.48**	**5.84**	**7.62**	**6.93**	**6.28**

E.2 REAL-TIME GDP FORECASTING

We also compare performance of B2F with the baselines for GDP forecasting task in Table 6.

Table 5: % improvement in MAE and MAPE scores averaged over all regions from Jan 2021 to June 2021

		k=1		k=2		k=3		k=4	
Cand. Model	Refining Model	**MAE**	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**	**MAE**	**MAPE**
ENSEMBLE	FFN	-0.26	-0.55	-0.35	-0.12	0.43	-0.29	0.87	0.77
	B2F-MB	-1.75	-0.74	-2.23	-1.57	-1.63	-3.66	-2.19	-2.85
	B2F-NOGRAPH	0.08	0.27	-1.45	-2.73	-1.42	-3.88	-5.72	-6.72
	B2F-BSEQ	1.85	0.28	1.42	0.37	0.65	0.28	0.74	0.44
	B2F	**6.31**	**6.04**	**5.25**	**4.39**	**3.31**	**4.19**	**4.41**	**3.15**
GT-DC	FFN	-2.66	-1.35	-2.42	-1.51	-1.49	-6.65	-1.54	-0.48
	B2F-MB	-3.51	-4.21	-3.02	-3.41	-2.72	-4.61	-2.91	-3.22
	B2F-NOGRAPH	0.26	0.67	2.24	3.51	2.34	2.54	1.93	0.78
	B2F-BSEQ	2.29	2.95	2.13	3.84	2.37	3.52	1.08	2.33
	B2F	**9.61**	**10.2**	**10.33**	**11.84**	**10.79**	**10.93**	**9.92**	**11.27**
YYG	FFN	-3.62	-1.53	-2.08	-1.34	-4.41	-3.88	-2.64	-3.36
	B2F-MB	-1.35	-1.69	-3.84	-6.99	-5.31	-4.6	-8.84	-5.61
	B2F-NOGRAPH	-1.74	-0.56	-1.25	-0.7	-2.85	-2.31	-6.13	-5.31
	B2F-BSEQ	-2.62	-1.84	-1.78	-2.26	-2.67	-1.17	-0.79	-0.62
	B2F	**9.52**	**6.14**	**8.93**	**6.32**	**6.82**	**7.58**	**7.32**	**5.73**
UMASS-MB	FFN	-2.56	-1.44	-3.25	-5.74	-1.46	-2.58	-1.01	-5.28
	B2F-MB	-3.88	-3.85	-8.2	-7.54	-5.55	-8.12	-6.49	-7.56
	B2F-NOGRAPH	-1.83	-0.91	-2.16	-1.88	-2.73	-0.76	-2.15	-2.87
	B2F-BSEQ	0.66	0.73	1.58	0.86	-1.12	0.36	0.36	0.96
	B2F	**4.51**	**4.75**	**5.43**	**4.66**	**5.68**	**4.89**	**3.32**	**3.11**
CMU-TS	FFN	-3.11	-4.46	-5.24	-4.93	-3.52	-2.32	-3.12	-0.65
	B2F-MB	-6.81	-5.92	-8.17	-8.21	-3.75	-7.6	-3.72	-6.11
	B2F-NOGRAPH	0.41	0.85	-0.67	-0.57	-0.73	-0.29	-0.46	-1.77
	B2F-BSEQ	1.33	2.46	1.46	1.05	2.17	2.17	2.38	2.26
	B2F	**7.54**	**8.22**	**7.5**	**8.04**	**6.42**	**8.23**	**5.73**	**6.22**

Table 6: % improvement in MAE and MAPE scores for GDP Forecasting from 2000 to 2021

		k=1		**k=2**	
Cand. Model	Refining Model	**MAE**	**MAPE**	**MAE**	**MAPE**
VAR	FFN	2.44±0.66	2.22 ± 0.40	3.09±0.39	4.23±0.24
	B2F-MB	4.05± 0.32	3.73±0.13	3.48± 0.23	3.04±0.29
	B2F-NOGRAPH	3.32±0.23	2.40±0.81	3.39±0.14	2.99±0.48
	B2F-BSEQ	6.36±0.38	6.11±0.15	5.54±0.42	4.95±0.23
	B2F	**15.88 ± 0.18**	**15.71±0.63**	**14.94±0.34**	**10.26±0.56**
FFN	FFN	-0.11±0.48	-0.15±0.74	0.48±0.83	0.36±0.50
	B2F-MB	0.28±0.44	-0.07±0.13	-0.24±0.48	-0.23±0.62
	B2F-NOGRAPH	2.26±0.43	1.98±0.85	2.29±0.18	2.20±0.36
	B2F-BSEQ	2.85±0.51	2.55±0.39	2.39±0.13	3.77±0.42
	B2F	**6.63 ± 0.13**	**7.30±0.33**	**6.07±0.44**	**6.61± 0.52**
GRU	FFN	-1.13 ± 0.03	0.82±0.26	-1.92±0.42	-0.70±0.21
	B2F-MB	0.83±0.25	0.65±0.38	0.99±0.37	1.02±0.26
	B2F-NOGRAPH	0.62±0.19	1.29±0.47	0.92±0.27	1.15±0.66
	B2F-BSEQ	2.15±0.66	3.04±0.73	3.78±0.58	2.51±0.61
	B2F	**7.52±0.40**	**7.18±0.27**	**7.28±0.69**	**7.59±0.88**