Appendix A: Ohio Dataset Experiments

Data and Training Description

This dataset includes both the OHIOT1DM 2018 and 2020 datasets, developed for the Knowledge Discovery in Healthcare Data Blood Glucose Level Predication Challenge (Marling and Bunescu 2020). The data pertain to 12 individuals, each with approximately 10,000 5-minute samples for training and 2,500 for testing, with carbohydrate administrations occurring every 88 timepoints on average, (median, [IQR]: 70, [56,134]), and insulin boluses occurring every 52 timepoints on average (36, [28,63]). 12% of glucose values are missing, but we do not include windows with missing glucose values.

This dataset contains the same variables and is processed and analyzed identically to the simulated dataset, except as described here. For the real dataset we evaluated on the held-out test data from the challenges. The remaining data were split into 80% train and 20% validation. Models were trained for at least 25 epochs, and then until validation data performance did not improve for 10 epochs.

The Ohio Dataset (Ohio T1D Blood Glucose Level Prediction Challenge, 2018 and 2020), can be made available through a data-use agreement with the owners: http://smarthealth.cs.ohio.edu/OhioT1DM-dataset.html.

Primary Results

On the Ohio dataset, performance gains are more moderate (rMSE 20.16 vs 20.36, **Table 2**), when compared to the simulated dataset. Multiple approaches exhibit negative SIV usage for the Ohio dataset, indicating that including the SIVs does more harm then good. We hypothesize that this is due to noise in the carbohydrate signal.

Table 2: Forecasting Error and SIV usage for the real dataset. Outcomes are reported as: Error [95% confidence interval] (SIV Usage). Our proposed approach outperforms baseline, although to a lesser degree than the simulated dataset. Confidence intervals were calculated from bootstraps with 1,000 resamples.

Model	rMSE [95%CI] (Usage)	MAE [95%CI] (Usage)
Encoder/Decoder	20.36,[19.46,21.30] (0.08)	14.67,[14.11,15.24] (0.24)
SIV Fine-tune	21.74,[20.87,22.64] (-1.30)	16.25,[15.68,16.85] (-1.35)
SIV Initialize	20.98,[20.00,21.97] (-0.54)	14.99,[14.42,15.59] (-0.09)
Full Capacity	20.98,[20.04,21.92] (-0.54)	15.09,[14.5,15.69] (-0.18)
Proposed	20.16,[19.28,21.06] (0.28)	14.64,[14.09,15.20] (0.27)

Individual Level Results and Ablations

With respect to the Ohio data, while the overall trend was the same as the simulated data, across individuals, the correlation between baseline error and our approach's improvement over baseline was not significant (r=0.22, p=0.49, **Figure 7 (b)**). We hypothesize that this again might be due to the presence of noise in the carbohydrate signal, which prohibits our model from accurately modeling the SIV signal (explored in Section 5.5). Alternatively, the intrinsic dynamics in the Ohio dataset may simply be more complex and thus result in more variability across individuals. The association between baseline encoder/decoder SIV usage and improvement over baseline does hold for real data (**Figure 7** (a)), Pearson r=-0.59, p=0.042,

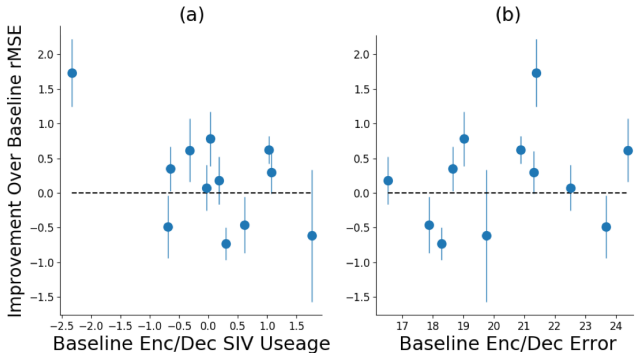

Figure 7: (a) Our architecture's improvement over the encoder/decoder baseline vs baseline SIV usage for the Ohio dataset. Our method's benefit increases as baseline SIV usage decreases. (b) Improvement over baseline vs baseline prediction error for Ohio data, for each individual. Improvement over baseline is not correlated with baseline error.

Table 3: rMSE and MAE, with SIV usage, for each ablation. Outcomes are reported as: Error [95% confidence interval] (SIV Usage). Confidence intervals were calculated from bootstraps with 1,000 resamples.

Model	rMSE [95%CI] (Usage)	MAE [95%CI] (Usage)
	Ohio	
No Gating	20.34,[19.48,21.27] (0.11)	14.77,[14.22,15.33] (0.13)
No Restriction	20.38,[19.48,21.31] (0.06)	14.73,[14.16,15.32] (0.18)
No Dec. SIV Input	20.71,[19.81,21.64] (-0.27)	14.97,[14.41,15.55] (-0.07)
Only Dec. SIV Input	20.52,[19.62,21.47] (-0.08)	14.70,[14.14,15.29] (0.20)
Proposed	20.16,[19.28,21.06] (0.28)	14.64,[14.09,15.20] (0.27)

The restriction element is important for the Ohio dataset (**Table 3**, rMSE increases to 20.38 from 20.16 when restriction is removed). This is likely because this dataset presents a more difficult challenge, compared to the simulated dataset, due to noise in the SIV signal and more complex target variable dynamics. For the Ohio dataset, we see a decrease in performance for each ablation. Our architecture works by isolating the effect that the SIV signal has on the target variable and enforcing consistency with domain knowledge. Although the domain knowledge is very general (we only restrict the signal direction), it improves performance, offering a benefit over isolation alone for the Ohio dataset. More restrictive model guidelines, such as directly restricting the architecture to use a detailed physiological model, could be beneficial, but during model development, we found that "less is more," in that a small amount of restriction with significant flexibility was most effective. However, some sort of domain-knowledge-based-guidance is helpful to overcome the challenges posed by the SIV problem, since without it, it is difficult to learn anything useful from the small number of non-zero samples.

Appendix B: Impact of carry-forward approach

Utilizing the Carry-forward approach improves performance on both datasets for both the baseline encoder/decoder and our proposed approach (**Table 4**).

Table 4: Forecasting Error and SIV usage for both datasets, examining our primary baseline and proposed approach with and without our carry-forward approach. Outcomes are reported as: Error [95% confidence interval] (SIV Usage). Both methods benefit from utilizing the carry-forward approach on both datasets. Confidence intervals were calculated from bootstraps with 1,000 resamples.

Model	rMSE [95%CI] (Usage)	MAE [95%CI] (Usage)
	Simulated- Carry Forward	
Encoder/Decoder	15.63,[14.08,16.89] (11.13)	12.42,[11.14,13.59] (6.63)
Proposed	**13.07,[11.77,14.16] (13.69)**	**10.45,[9.37,11.37] (8.61)**
	Simulated- NO Carry Forward	
Encoder/Decoder	16.46,[14.64,17.84] (10.30)	12.97,[11.53,14.13] (6.09)
Proposed	16.08,[14.46,17.37] (10.68)	12.80,[11.43,13.91] (6.25)
	Ohio- Carry Forward	
Encoder/Decoder	20.36,[19.46,21.30] (0.08)	14.67,[14.11,15.24] (0.24)
Proposed	**20.16,[19.28,21.06] (0.28)**	**14.64,[14.09,15.20] (0.27)**
	Ohio- NO Carry Forward	
Encoder/Decoder	20.64,[19.74,21.56] (-0.20)	14.98,[14.43,15.56] (-0.08)
Proposed	20.41,[19.53,21.35] (0.03)	14.85,[14.28,15.41] (0.05)

Table 5: Proportion of points in each region of the Clarke Error Grid. Region A represents strong forecasts, while regions C through E represent potentially catastrophic errors

Region	Baseline	Proposed
A	97.1	97.8
B	2.36	1.43
C	0.00	0.00
D	0.05	0.08
E	0.00	0.00

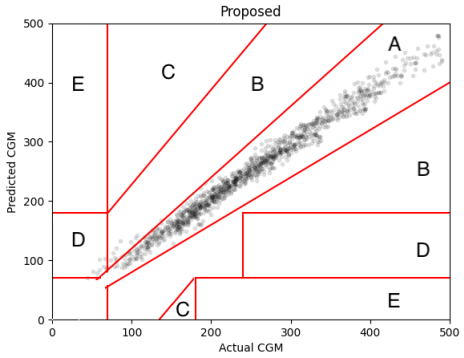

Figure 8: Clarke error grid for the proposed approach.

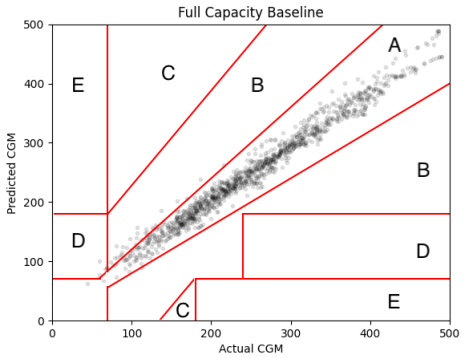

Figure 9: Clarke error grid for the full capacity baseline.

Appendix C: Clarke Error Grid

A Clarke error grid demonstrates where a forecaster could lead to catastrophic failure; predicted BG values are compared to true values, and regions where making treatment decisions based on forecasts would lead to poor health outcomes are highlighted. Clarke error grids for our approach and the best performing baseline on the simulated dataset, are shown in **Figure 8** and **Figure 9**, respectively. Both approaches demonstrate fairly strong performance, but our approach has 98% of points in region A, while the baseline has 97% of points in region A (**Table 5**). Region A represents the region where utilizing the forecasts for BG control would reliably lead to good health outcomes. While one percentage point is a modest improvement, every prediction is important in a clinical setting. This illustrates that while our approach is not a complete solution to reliable BG forecasting, it a step in the right direction.