Regimen-Aware Forecasting for Mechanistic Virtual Patients with Time-Series Foundation Models

Charles George Barker*, **Ahmad Wisnu Mulyadi***, **Sanjana Balaji Kuttae, Ansh Kumar, Lilija Wehling, Thomas Rückle, Gurdeep Singh**†
Virtual Patient Engine, BioMed X Germany GmbH, Heidelberg, Germany

Sommer Anjum[2], **Anastasios Siokis**[2], **Hans-Christoph Schneider**[2],
Sven Jager[1], **Thomas Klabunde**[2], **Tommaso Andreani**[2†]
[1]Digital R&D, Data & Computational Science, Sanofi Digital, Frankfurt am Main, Germany
[2]Translational Medicine Unit (TMU) - Disease Modeling, Sanofi R&D, Frankfurt am Main, Germany

Abstract

Quantitative Systems Pharmacology (QSP) models provide mechanistic representations of disease progression and drug response in virtual patient populations, but their construction and maintenance are resource-intensive. Large pretrained probabilistic time-series foundation models (TSFMs) exhibit strong cross-domain generalization, yet their application in mechanistically informed drug development remains limited. Here, we present a framework that augments QSP-based virtual patient generation with a fine-tuned TSFM surrogate. We fine-tuned Chronos-2 on a synthetic dataset generated from a curated portfolio of mechanistic QSP models and evaluated its performance on an unseen inflammatory bowel disease (IBD) QSP model. Forecasts were informed with explicit context in the form of historical and future drug dosing trajectories as exogenous inputs. Regimen-aware fine-tuning reduced predictive error and tightened uncertainty estimates relative to zero-shot inference. Incorporating dosing context alongside fine-tuning lowered both mean prediction error and variability across virtual patients. These results demonstrate that context-informed TSFMs can serve as scalable, reproducible surrogates for QSP workflows, enabling rapid scenario exploration in pharmacological modeling pipelines.

Track: Research

1 Introduction

Quantitative Systems Pharmacology (QSP) models provide mechanistically grounded, time-resolved descriptions of disease biology and drug response, and are increasingly used to support regulatory decision-making (Bai et al., 2024). Virtual "digital twin" from QSP simulations have been applied to predict efficacy prior to human trials (Klabunde, 2024), particularly in settings such as rare diseases where patient data are limited (Neves-Zaph & Kaddi, 2024). In addition, *in silico* trial emulation using mechanistic cardiovascular models demonstrates how synthetic virtual cohorts can be leveraged to predict long-term clinical outcomes and de-risk development programs (Angoulvant et al., 2024). At the same time, QSP development and maintenance require substantial domain expertise and iterative curation as biological knowledge evolves. As model scope and mechanistic detail grow, simulation and ensemble generation can become computationally intensive, particularly when scaling to large virtual cohorts or exploring many candidate dosing regimens. These practical constraints can limit reuse and rapid iteration, especially in an era where high-dimensional molecular and clinical data are becoming more widely available.

*These authors contributed equally to this work.

†Correspondence: gsingh@bmedx.com; tommaso.andreani@sanofi.com

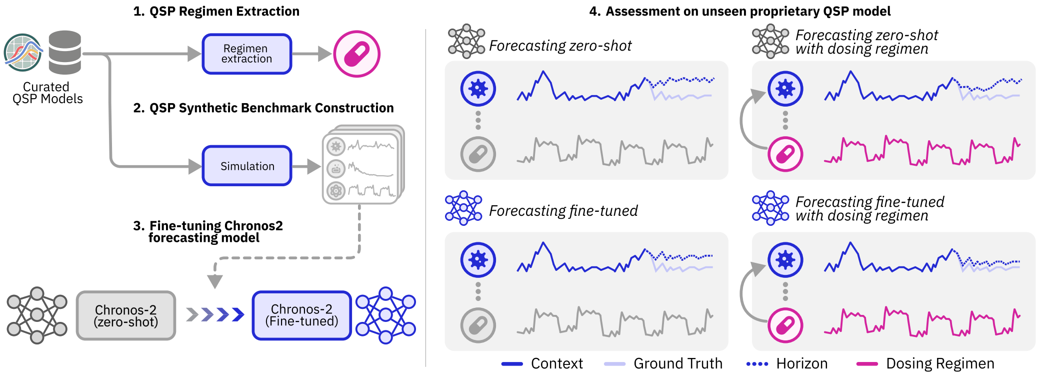
1. QSP Regimen Extraction
Curated QSP Models
Regimen extraction
2. QSP Synthetic Benchmark Construction
Simulation
3. Fine-tuning Chronos2 forecasting model
Chronos-2 (zero-shot)
Chronos-2 (Fine-tuned)
4. Assessment on unseen proprietary QSP model
Forecasting zero-shot
Forecasting zero-shot with dosing regimen
Forecasting fine-tuned
Forecasting fine-tuned with dosing regimen
Context
Ground Truth
Horizon
Dosing Regimen

Figure 1: Schematic overview of the framework outlined in this work.

In parallel, large pretrained time-series foundation models (TSFMs) have demonstrated strong zero-shot generalisation across heterogeneous domains, offering standardised architectures and efficient inference suitable for large-scale deployment (Auer et al., 2025; Ansari et al., 2025; Liu et al., 2025). Recent work has begun to explore whether such models can serve as zero-shot surrogates for simulating QSP-based virtual patient (VP) trajectories, highlighting their potential to capture pharmacodynamic patterns (Mulyadi et al., 2025a). Advances in context-aware time-series modelling enable forecasts to be conditioned not only on historical observations, but also on covariates that describe system attributes and interventions (Ansari et al., 2025). In medicine, time-series models have largely centred on electronic health records rather than mechanistically structured, multivariate trajectories typical of pharmacology (Shmatko et al., 2025; Li et al., 2025). This creates an opportunity to bridge mechanistic QSP modeling with flexible, data-driven TSFMs that can learn to incorporate intervention context. By combining mechanistic QSP models with TSFMs, we can leverage the interpretability and pharmacological grounding of QSP alongside the scalability and compositional learning capabilities of modern TSFMs.

Here, we present a framework for fine-tuning a TSFM on QSP-generated virtual patient trajectories with explicit dosing regimen conditioning (Figure 1). We selected Chronos-2 (Ansari et al., 2025) as the TSFM for this study because it outperforms comparable models with the lowest MASE values on GIFT-Eval (Aksu et al., 2024)[1], while supporting fine-tuning and zero-shot forecasting for diverse datasets. The original Chronos model consistently performed well in a prior benchmarking study across multiple biological and QSP-generated time-series trajectories, where it outperformed comparable TSFMs in forecasting accuracy(Mulyadi et al., 2025a). Using synthetic time-series data simulated from a suite of curated mechanistic QSP models, we fine-tuned Chronos-2 to predict clinically relevant species of an inflammatory bowel disease (IBD) QSP model while conditioning on dosing regimens as exogenous covariates. We demonstrate that incorporating dosing regimen information enhances predictive performance compared to models that disregard intervention context, and that fine-tuning on QSP-derived trajectories reduces predictive uncertainty compared to zero-shot use of the foundation model.

2 Proposed Method

QSP-derived Simulated Data We simulate a QSP model to obtain time-series data $\{\mathbf{X}^r\}_{r=1}^{|\mathcal{R}|}$, where $\mathbf{X}^r \in \mathbb{R}^{P \times S \times T}$ denotes a cohort of VPs treated under a regimen $r \in \mathcal{R}$, with P patients, covering S species over T timesteps. Hereafter, we omit indices for clarity. We further decompose the time-series into $\mathbf{X}_{1:T} = (\mathbf{X}^c, \mathbf{X}^h)$, where $\mathbf{X}^c = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{t=c}]$ and $\mathbf{X}^h = [\mathbf{X}_{t+1}, \mathbf{X}_{t+2}, \ldots, \mathbf{X}_{T=c+h}]$ denote the context and horizon segments, respectively. We feed $\mathbf{X}^c$ to the forecasting model, and use $\mathbf{X}^h$ as the ground-truth segment to compare against the predicted horizon $\widetilde{\mathbf{X}}^h$ when evaluating forecasting performance.

[1] last accessed on February 13, 2026; https://huggingface.co/spaces/Salesforce/GIFT-Eval

QSP-derived Regimen Course For each QSP model, we identified the species which represents the drugs effect (or the regimen $\mathcal{R}$) in the model and provided this to a TSFM (*i.e.*, Chronos-2) as an exogenous variable $\phi \subset S$. In QSP models, dosing species are the explicit state variables in which dose events are applied and accumulated; they therefore serve as complete proxies for the dosing trajectory over time, capturing timing, magnitude, and route of administration without requiring additional mechanistic states. We incorporated them as additional covariates for which both context and horizon $\mathbf{X}^c$ and $\mathbf{X}^h$ are given. The rationale for using the species as a regimen-proxy are found in Table 4 in the Appendix.

Conditional Time-series Forecasting Since QSP models encode drug-dosing regimens as dedicated species, we can mimic clinical settings in which treatment is pre-specified and observed throughout the entire course. We formalize the conditional forecasting task as predicting a horizon segment, $\widetilde{\mathbf{X}}^h = f_\theta(\mathbf{X}^c \mid \mathbf{X}^\phi)$, where f_θ denotes a TSFM with pre-trained parameters θ, and $\mathbf{X}^\phi$ denotes time-series of exogenous variable(s). In practice, we optimize θ via fine-tuning.

3 Experimental Results

Implementation Details We curated 12 total number of QSP-simulated data inferred from publicly-available BioModels database (Malik-Sheriff et al., 2020). For every model, simulation time-courses were extracted from the corresponding publications and reproduced using COPASI version 4.44 (build 295). All models were executed under their reported experimental conditions to generate time-series data. We applied global z-score normalization across all VPs, treatments, and time points (according to Eq. 1). We excluded species with zero variance, as these were invariant across all simulations and uninformative for downstream inference. In total, we obtained $304,971$ time-series samples spanning models, treatments, and VPs. Furthermore, we utilized 25% of the total timesteps from each sample as the context for the model. We fine-tuned Chronos-2 using a LoRA adapter with a 0.9:0.1 train–validation split and evaluated generalization on an unseen proprietary IBD QSP model. Fine-tuning was performed with a batch size of 32, learning rate of $1e-6$, for $20,000$ steps on a single NVIDIA H100 80GB GPU. Forecasting performance was evaluated using MSE, RMSE, MAE, and SMAPE. As Chronos-2 produces stochastic forecasts, we additionally assessed predictive uncertainty using prediction interval coverage probability (PICP) and mean prediction interval width (MPIW) (Pearce et al., 2018). See Appendix C for details.

Regimen Conditioning Improves Predictive Accuracy Table 1 presents the primary benchmarking results, reporting predictive performance across held-out trajectories of an IBD QSP model under different context strategies. Including the dosing regimen as an exogenous context substantially improves performance, reducing both error and variability across virtual patients compared to models without regimen information. The relatively high variance arises from parameter-diverse mechanistic simulations of virtual patients, reflecting the heterogeneity of real patient backgrounds. To ameliorate this variance, we also assessed treatment-level inferences, which show that for therapies with regular, fixed schedules (*e.g.*, Nanobody(R)-based proprietary compound), forecasts remain accurate even in zero-shot settings without explicit dosing context, reflecting the predictability of the administration pattern. In contrast, for treatments with irregular or adaptive dosing regimens (*e.g.*, Guselkumab), incorporating dosing trajectories as contextual input markedly enhances predictive accuracy (Figure 2). This holds true across species (Appendix Figure 3) and also for calprotectin (Appendix Figure 4), a diagnostic protein for IBD.

Fine-Tuning Improves Predictive Uncertainty Fine-tuning substantially reduces predictive uncertainty. We quantified uncertainty using the PICP and MPIW (Table 2), Figure 5). In the zero-shot setting, PICP is 0.73 and MPIW is 0.44, indicating wide, over-conservative prediction intervals. After fine-tuning on the synthetic QSP dataset, PICP decreases to approximately 0.60 while MPIW contracts to 0.2, demonstrating sharper predictions. This reduction in both coverage and interval width indicates that exposure to mechanistically generated QSP trajectories allows the TSFM to produce sharper, more confident predictions.

Treatment-specific effects in the unseen IBD model To assess whether regimen conditioning improves performance consistently across therapies, we performed paired Wilcoxon signed-rank

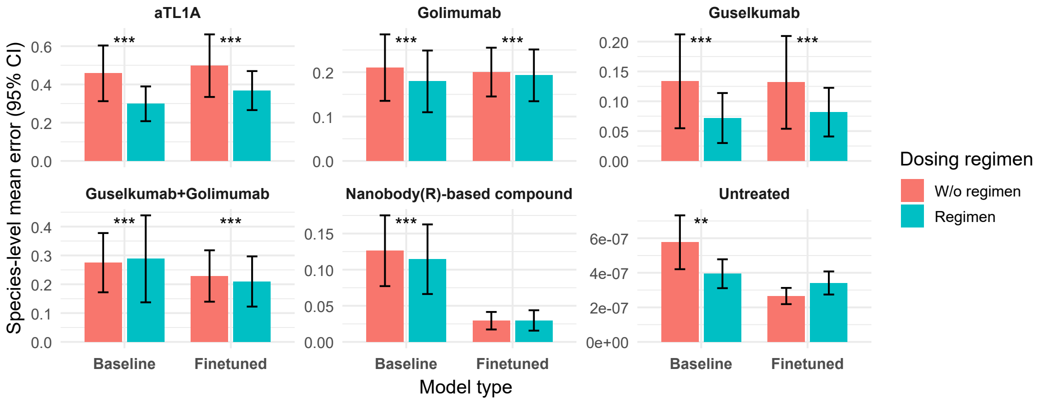

Figure 2: Species-level mean RMSE (bars) with 95% confidence intervals across treatments, aggregated over all virtual patients, comparing models trained with versus without dosing regimen information. Results are shown separately for baseline and finetuned models. Significance markers indicate paired Wilcoxon tests across species (BH-corrected), testing whether inclusion of dosing regimen reduces error.

tests comparing forecasting error with and without regimen covariates at the species level (158 paired species per treatment, Table 3). Multiple testing correction was performed using Benjamini–Hochberg adjustment. Across treatments, we see statistically significant error reductions for treatments (small–medium effect sizes, BH-adjusted $p < 0.001$), and negligible effect for untreated patients, who already have a far lower baseline error (Table 3).

Case study: Gut Calprotectin under Guselkumab treatment We next examined a clinically relevant biomarker in IBD: Gut Calprotectin levels under Guselkumab (Gusel) treatment (Appendix Figure 6). Calprotectin is a validated biomarker of intestinal inflammation and is widely used for disease monitoring in IBD (Kapel et al., 2023; Plevris & Lees, 2022). For this biomarker, inclusion of dosing regimen trajectories as exogenous covariates led to qualitatively improved forecasts. Compared to forecasts generated without regimen context, the conditioned model more accurately reflected the expected pharmacodynamic response following drug administration.

4 Limitations and Future Works

Despite promising results, several limitations of the current framework remain. First, treatment regimens are encoded as simplified temporal inputs, whereas real-world dosing corresponds to complex PK–PD trajectories that vary across patients and over time. Accurately capturing these dynamics without access to full concentration–time data remains challenging, particularly in clinical settings where PK sampling is sparse or unavailable. Second, the model requires sufficient longitudinal history to forecast trajectories, but patient data are often sparse or limited to baseline measurements. Extending the framework to incorporate snapshot modalities, such as baseline biomarkers or omics profiles, using multimodal architectures that translate static patient information into dynamic time-series predictions represents a promising direction (Arango et al., 2025; Zhong et al., 2025; Kimura et al., 2025). For instance, a TSFM combined with a conditional autoencoder could leverage task-specific multimodal inputs and contrastive learning to generate informative latent representations that guide temporal dynamics (Qiu et al., 2025).

Third, while the model achieves improved predictive accuracy, it does not provide explicit causal or mechanistic explanations for its forecasts, which is often required in translational and regulatory contexts. Integrating reasoning frameworks or structured biomedical knowledge (e.g., (Mulyadi et al., 2025b)) could improve transparency and enhance regulatory relevance. The compositional nature of TSFMs provides a natural pathway for integrating patient-specific, multimodal data, enabling more personalized and extensible modeling frameworks beyond what is feasible with purely mechanistic QSP models. While the present study focuses on regimen-aware forecasting from QSP-generated trajectories, it establishes a foundation for future approaches that incorporate heterogeneous clinical and molecular data into a unified, intervention-aware modeling framework. Overall, these limitations highlight key opportunities to refine regimen encoding, expand contextual representations, and increase mechanistic interpretability, laying the groundwork for more physiologically accurate context-aware TSFMs. These developments will further extend the applicability of the framework across a broader range of mechanistic disease models, and improve robustness of regimen-aware forecasting under diverse biological settings.

5 Conclusions

We introduce a regimen-aware adaptation of time-series foundation models by fine-tuning Chronos-2 on a large synthetic dataset generated from diverse QSP models. Conditioning forecasts on dosing trajectories improves accuracy and reduces uncertainty on an unseen IBD model relative to zero-shot inference, demonstrating that TSFMs can serve as scalable surrogates for virtual patient forecasting. This effect is not observed in untreated simulations, where baseline errors are substantially lower and changes in RMSE are negligible in absolute terms, reflecting a regime in which the regimen input corresponds to a constant no-intervention signal and provides no additional informative conditioning. Although evaluation is performed on an IBD QSP model, it represents an unseen system relative to training, and our prior work has shown that TSFMs generalize across multiple biological contexts(Mulyadi et al., 2025a); notably, IBD has been identified as a particularly challenging system, and our method improves performance in this difficult setting. While our current approach does not leverage mechanistic structure beyond the simulated data itself, our results suggest a promising path toward integrating richer mechanistic context and multimodal biological information in future time-series foundation models, enabling more robust intervention-aware forecasting in biomedical domains.

Acknowledgments

We acknowledge Julia Deichmann for her in-depth discussions and help with the IBD proprietary model from Sanofi. We also acknowledge Markus Rehberg for his feedback on the manuscript.

Funding

This study was funded by Sanofi.

Conflict Interests

SA, AS, HCS, SJ, TK and TA are Sanofi employees and may hold shares and/or stock options in the company.

References

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. GIFT-Eval: A Benchmark For General Time Series Forecasting Model Evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Denis Angoulvant, Solène Granjeon-Noriot, Pierre Amarenco, Alexandre Bastien, Emmanuelle Bechet, Franck Boccara, Jean-Pierre Boissel, Bertrand Cariou, Eulalie Courcelles, Alizée Diatchenko, et al. In-silico trial emulation to predict the cardiovascular protection of new lipid-lowering drugs: an illustration through the design of the SIRIUS programme. *European Journal of Preventive Cardiology*, 31(15):1820–1830, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Sebastian Pineda Arango, Pedro Mercado, Shubham Kapoor, Abdul Fatir Ansari, Lorenzo Stella, Huibin Shen, Hugo Senetaire, Caner Turkmen, Oleksandr Shchur, Danielle C Maddix, et al. ChronosX: Adapting pretrained time series models with exogenous variables. *arXiv preprint arXiv:2503.12107*, 2025.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. TiRex: Zero-Shot Forecasting Across Long and Short Horizons with Enhanced In-Context Learning. *arXiv preprint arXiv:2505.23719*, 2025.

Jane PF Bai, Guansheng Liu, Miao Zhao, Jie Wang, Ye Xiong, Tien Truong, Justin C Earp, Yuching Yang, Jiang Liu, Hao Zhu, et al. Landscape of regulatory quantitative systems pharmacology submissions to the US Food and Drug Administration: An update report. *CPT: Pharmacometrics & Systems Pharmacology*, 13(12):2102–2110, 2024.

N Benson, E Metelkin, O Demin, GL Li, D Nichols, and PH Van Der Graaf. A systems pharmacology perspective on the clinical development of fatty acid amide hydrolase inhibitors for pain. *CPT: pharmacometrics & systems pharmacology*, 3(1):1–7, 2014.

Lisette G de Pillis, K Renee Fister, Weiqing Gu, Tiffany Head, Kenny Maples, Todd Neal, Anand Murugan, and Kenji Kozai. Optimal control of mixed immunotherapy and chemotherapy of tumors. *Journal of Biological systems*, 16(01):51–80, 2008.

G Dwivedi, L Fitz, M Hegen, SW Martin, J Harrold, A Heatherington, and C Li. A Multiscale Model of Interleukin-6–Mediated Immune Regulation in Crohn's Disease and Its Application in Drug Discovery and Development. *CPT: Pharmacometrics & Systems Pharmacology*, 3(1):1–9, 2014.

Dana Faratian, Alexey Goltsov, Galina Lebedeva, Anatoly Sorokin, Stuart Moodie, Peter Mullen, Charlene Kay, In Hwa Um, Simon Langdon, Igor Goryanin, et al. Systems biology reveals new strategies for personalizing cancer medicine and confirms the role of PTEN in resistance to trastuzumab. *Cancer research*, 69(16):6713–6720, 2009.

Kapil Gadkar, Christina Friedrich, Vincent Hurez, Maria-Luisa Ruiz, Leslie Dickmann, Mohit Kumar Jolly, Leah Schutt, Jin Jin, Joseph A Ware, and Saroja Ramanujan. Quantitative systems pharmacology model-based investigation of adverse gastrointestinal events associated with prolonged treatment with PI3-kinase inhibitors. *CPT: Pharmacometrics & Systems Pharmacology*, 11(5):616–627, 2022.

Nathalie Kapel, Hamza Ouni, Nacer Adam Benahmed, and Laurence Barbot-Trystram. Fecal calprotectin for the diagnosis and management of inflammatory bowel diseases. *Clinical and Translational Gastroenterology*, 14(9):e00617, 2023.

Tomoyoshi Kimura, Xinlin Li, Osama Hanna, Yatong Chen, Yizhuo Chen, Denizhan Kara, Tianshi Wang, Jinyang Li, Xiaomin Ouyang, Shengzhong Liu, et al. Infomae: Pair-efficient cross-modal alignment for multimodal time-series sensing signals. In *Proceedings of the ACM on Web Conference 2025*, pp. 3084–3095, 2025.

Thomas Klabunde. Digital "Twinning": Clinical Trials Powered by AI. `https://www.sanofi.com/en/magazine/our-science/digital-twinning-clinical-trials-ai`, 2024. Accessed: 2026-02-13.

Odelaisy León-Triana, Antonio Pérez-Martínez, Manuel Ramírez-Orellana, and Víctor M Pérez-García. Dual-target CAR-Ts with on-and off-tumour activity may override immune suppression in solid cancers: A mathematical proof of concept. *Cancers*, 13(4):703, 2021a.

Odelaisy León-Triana, Soukaina Sabir, Gabriel F Calvo, Juan Belmonte-Beitia, Salvador Chulián, Álvaro Martínez-Rubio, María Rosa, Antonio Pérez-Martínez, Manuel Ramirez-Orellana, and Víctor M Pérez-García. CAR T cell therapy in B-cell acute lymphoblastic leukaemia: Insights from mathematical models. *Communications in Nonlinear Science and Numerical Simulation*, 94:105570, 2021b.

Hao Li, Bowen Deng, Chang Xu, ZhiYuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, and Jiang Bian. MIRA: Medical Time Series Foundation Model for Real-World Health Data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025.

Rahuman S Malik-Sheriff, Mihai Glont, Tung VN Nguyen, Krishna Tiwari, Matthew G Roberts, Ashley Xavier, Manh T Vu, Jinghao Men, Matthieu Maire, Sarubini Kananathan, et al. Biomodels—15 years of sharing computational models in life science. *Nucleic acids research*, 48(D1): D407–D415, 2020.

Ahmad Wisnu Mulyadi, Charlie George Barker, Sanjana Balaji Kuttae, Lilija Wehling, Thomas Rückle, Nicolas Boucher, Firas Abdessalem, Sven Jager, Anastasios Siokis, Sommer Anjum, Mohammed H. Mosa, Thomas Klabunde, Tommaso Andreani, and Gurdeep Singh. Evaluating Time-Series Foundation Models as Zero-Shot Surrogates for Mechanistic Virtual Patients. In *EurIPS 2025 Workshop on SIMBIOCHEM*, 2025a.

Ahmad Wisnu Mulyadi, Lilija Wehling, Ansh Kumar, Nicolas Boucher, Firas Abdessalem, Sven Jager, Mohammed H. Mosa, Thomas Klabunde, Tommaso Andreani, and Gurdeep Singh. BioMedReasoner: Towards Multi-Hop Reasoning using Path-based Relational Learning on Biomedical Knowledge Graphs. In *NeurIPS 2025 AI for Science Workshop*, 2025b.

Susana Neves-Zaph and Chanchala Kaddi. Quantitative Systems Pharmacology Models: Potential Tools for Advancing Drug Development for Rare Diseases. *Clinical Pharmacology & Therapeutics*, 116(6):1442–1451, 2024.

Robert Palmér, Elin Nyman, Mark Penney, Anna Marley, Gunnar Cedersund, and Balaji Agoram. Effects of IL-1β–blocking therapies in type 2 diabetes mellitus: a quantitative systems pharmacology modeling approach to explore underlying mechanisms. *CPT: Pharmacometrics & Systems Pharmacology*, 3(6):1–8, 2014.

Tim Pearce, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4075–4084, 2018.

Nikolas Plevris and Charlie W Lees. Disease Monitoring in Inflammatory Bowel Disease: Evolving Principles and Possibilities. *Gastroenterology*, 162(5):1456–1475, 2022.

Jiaxing Qiu, Dongliang Guo, Brynne Sullivan, Teague R Henry, and Thomas Hartvigsen. Instruction-based Time Series Editing. *arXiv preprint arXiv:2508.01504*, 2025.

Johannes Schropp, Antari Khot, Dhaval K Shah, and Gilbert Koch. Target-Mediated Drug Disposition Model for Bispecific Antibodies: Properties, Approximation, and Optimal Dosing Strategy. *CPT: Pharmacometrics & Systems Pharmacology*, 8(3):177–187, 2019.

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256, 2025.

James P Sluka, Xiao Fu, Maciej Swat, Julio M Belmonte, Alin Cosmanescu, Sherry G Clendenon, John F Wambaugh, and James A Glazier. A liver-centric multiscale modeling framework for xenobiotics. *PLOS ONE*, 11(9):e0162428, 2016.

T Wajima, G K Isbister, and S B Duffull. A Comprehensive Model for the Humoral Coagulation Network in Humans. *Clinical Pharmacology & Therapeutics*, 86(3):290–298, 2009.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. *arXiv preprint arXiv:2502.04395*, 2025.

A Supplementary Figures

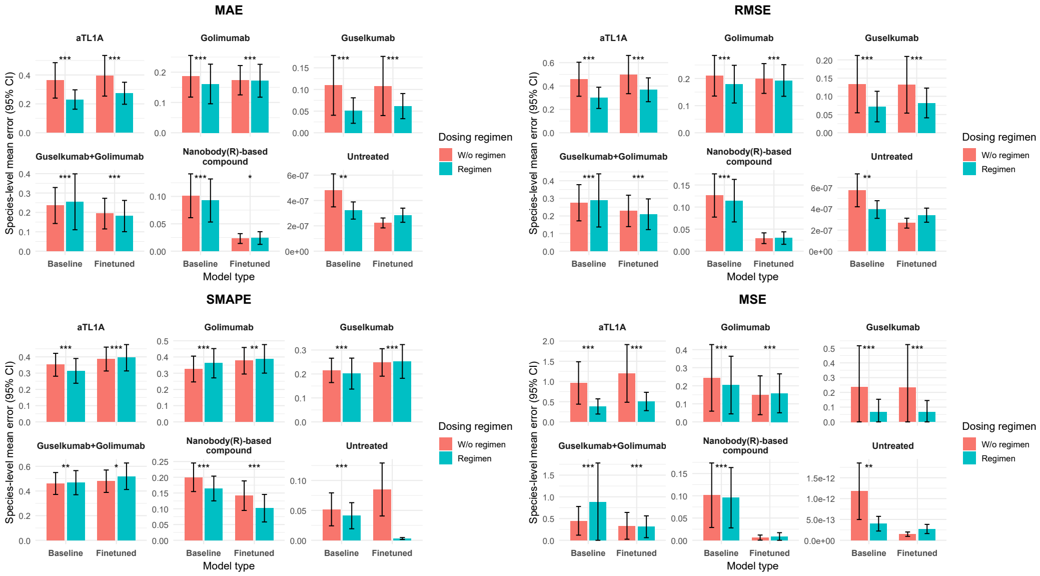

Figure 3: Species-level mean (bars) with 95% confidence intervals across treatments, comparing models trained with versus without dosing regimen information for different calculations of error. Results are shown separately for baseline and finetuned models. Significance markers indicate paired Wilcoxon tests across species (BH-corrected), testing whether inclusion of dosing regimen reduces error.

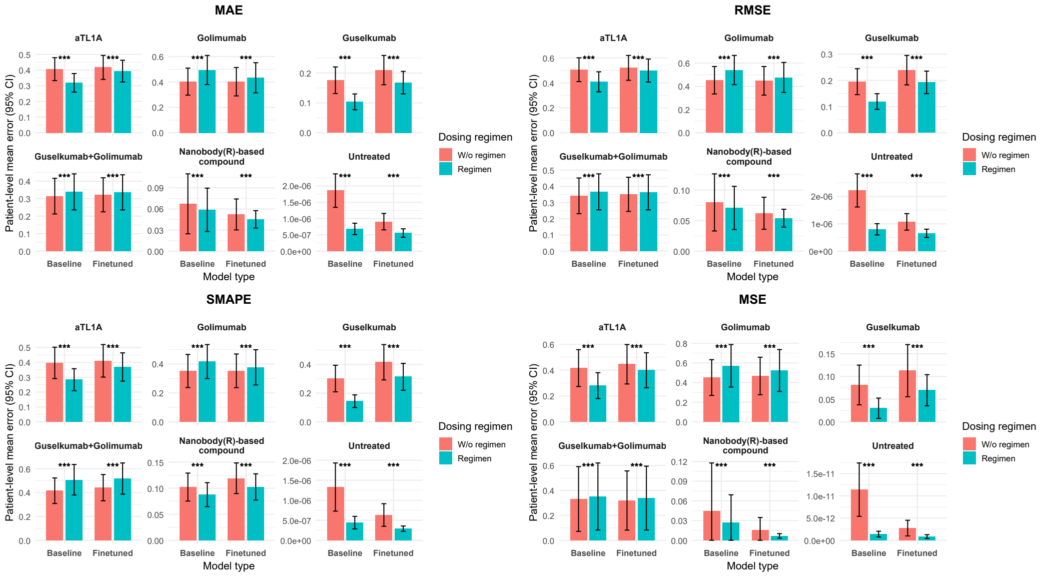

Figure 4: Mean error for predictions for gut calprotectin levels (bars) with 95% confidence intervals across treatments, comparing models trained with versus without dosing regimen information for different calculations of error. Results are shown separately for baseline and finetuned models. Significance markers indicate paired Wilcoxon tests across species (BH-corrected), testing whether inclusion of dosing regimen reduces error.

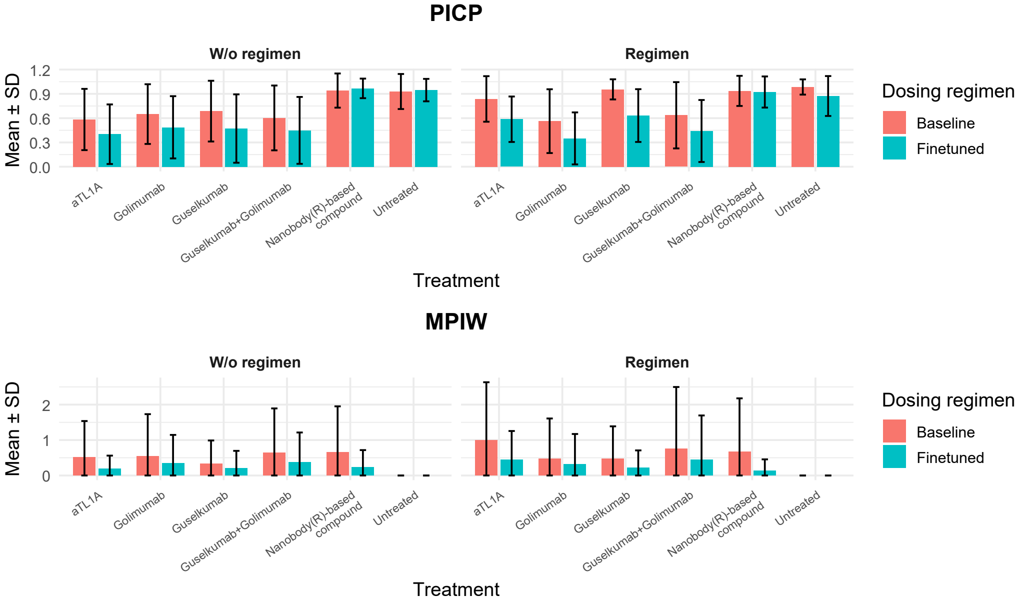

Figure 5: Mean prediction interval coverage probability (PICP, top) and mean prediction interval width (MPIW, bottom) across treatments for baseline and finetuned models, with and without dosing regimen information. Bars show mean values and error bars indicate ±1 standard deviation; metrics are aggregated under the no-covariates setting, and y-axes are scaled independently for each metric.

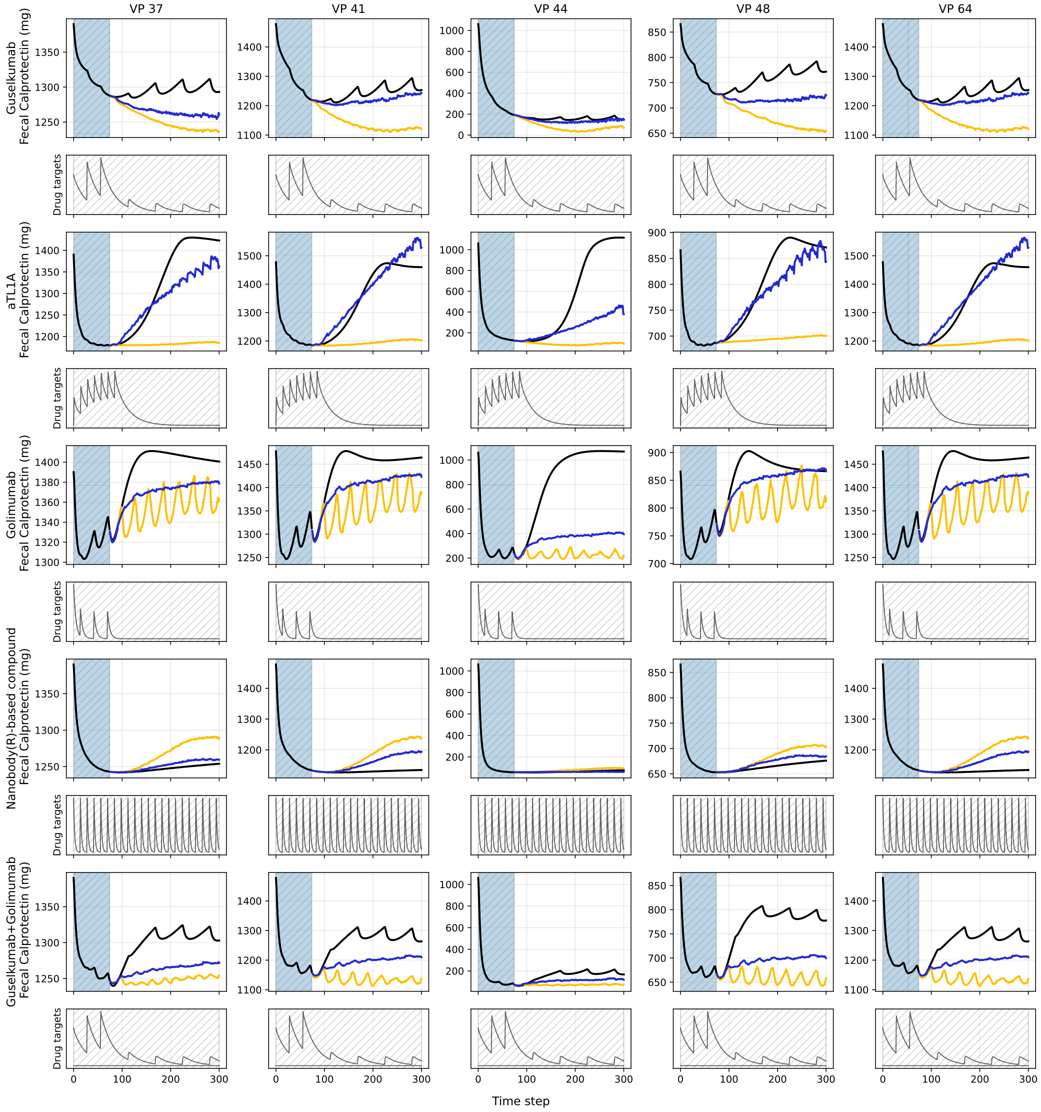

Figure 6: Simulated fecal calprotectin levels (y-axis) over time (x-axis) for selected virtual patients (labelled at VPs, as columns) across treatments different treatments (rows). Solid black lines denote observed context and ground-truth trajectories. Yellow lines show model forecasts without dosing variables, while blue lines show forecasts with dosing variables. Hatched blue regions indicate the context window used for prediction. Lower panels display the dynamics of drug-target species associated with each treatment.

B Supplementary Tables

Table 1: Predictive error metrics across training and regimen conditions. Mean ± standard deviation values are reported for MAE, MSE, RMSE, and SMAPE across virtual patients. Results are stratified by model training paradigm (Baseline vs. Finetuned) and inclusion of dosing regimen information (w/o regimen vs. w/ regimen).

Training	**Dosing Variable**	**MAE**	**MSE**	**RMSE**	**SMAPE**
Baseline	W/o regimen	0.229 ± 0.556	0.493 ± 2.537	0.275 ± 0.646	0.298 ± 0.436
	Regimen	0.190 ± 0.538	0.426 ± 3.606	0.228 ± 0.611	**0.285 ± 0.458**
Finetuned	W/o regimen	0.212 ± 0.544	0.465 ± 2.320	0.256 ± 0.632	0.317 ± 0.459
	Regimen	**0.178 ± 0.428**	**0.307 ± 1.348**	**0.217 ± 0.510**	0.305 ± 0.494

Table 2: Uncertainty quantification metrics across training and regimen conditions. Mean ± standard deviation values are reported for PICP and MPIW across virtual patients. Results are stratified by model training paradigm (Baseline vs. Finetuned) and inclusion of dosing regimen information (W/o regimen vs. Regimen).

Training	**Dosing Variable**	**PICP**	**MPIW**
Baseline	W/o regimen	0.732 ± 0.366	0.447 ± 1.038
	Regimen	0.870 ± 0.269	0.632 ± 1.607
Finetuned	W/o regimen	0.620 ± 0.411	0.224 ± 0.585
	Regimen	0.660 ± 0.353	0.296 ± 0.914

Table 3: Statistical analysis (Wilcoxon signed-rank tests) of RMSE error and change after adding regimen covariates. Multiple-hypothesis correction was performed using Benjamini-Hochberg adjustment. Direction column indicates whether errors have reduced or increased since taking into account regimens.

Training	**Treatment**	$p-$**value**	$p-$**adj**	Δ **RMSE**	**Direction**
Baseline	Nanobody(R)-based compound	1.999e-7	3.411e-7	-0.003	Reduced
	Golimumab	4.001-13	1.460-12	-0.015	Reduced
	Guselkumab+Golimumab	3.009-7	4.524-7	-0.019	Reduced
	Guselkumab	7.616-19	9.139-18	-0.0153	Reduced
	aTL1A	1.1863-10	4.472-10	-0.043	Reduced
	Untreated	0.003	0.003	1.032-8	Increased
Finetuned	Nanobody(R)-based compound	0.050	0.054	-9.471-5	Reduced
	Golimumab	1.708-4	2.277-4	-0.015	Reduced
	Guselkumab+Golimumab	3.433-10	6.866-10	-0.0176	Reduced
	Guselkumab	2.146-17	1.288-16	-0.009	Reduced
	aTL1A	4.868-13	1.460-12	-0.026	Reduced
	Untreated	0.999	0.999	1.949-8	Increased

Table 4: Summary of QSP models used for benchmark with reference and model details.

#	Reference	Treatment(s)	Virtual Patient(s)	Species	Parameters
1	Proprietary Model	6	71	183	829
2	(Gadkar et al., 2022)	8	704	59	188
3	(Faratian et al., 2009)	2	1	55	114
4	(Benson et al., 2014)	7	1	39	155
5	(Wajima et al., 2009)	1	1	54	58
6	(Palmér et al., 2014)	3	1	35	52
7	(Schropp et al., 2019)	4	1	8	21
8	(Dwivedi et al., 2014)	3	1	44	50
9	(Sluka et al., 2016)	1	1	7	9
10	(de Pillis et al., 2008)	1	1	6	31
11	(León-Triana et al., 2021b)	1	1	3	5
12	(León-Triana et al., 2021b)	3	1	4	10
13	(León-Triana et al., 2021a)	1	3	5	12

Table 5: Summary of QSP models used for benchmark. Each entry lists reference, simulation settings, and BioModels identifiers where available, respectively.

#	Reference	Simulation Time	Disease	BioModels ID
1	Proprietary Model	300 days	IBD	–
2	(Gadkar et al., 2022)	5760 hrs	Adverse GI events (intracellular/ extracellular dynamics in gut epithelia)	–
3	(Faratian et al., 2009)	60 mins	Cancer (HER2-positive breast cancer)	BIOMD0000000424
4	(Benson et al., 2014)	350 hrs	Osteoarthritic Pain	BIOMD0000000512
5	(Wajima et al., 2009)	20 days	Thrombotic disorders	BIOMD0000000340
6	(Palmér et al., 2014)	350 days	Type 2 Diabetes Mellitus	BIOMD0000000620
7	(Schropp et al., 2019)	150 days	Bispecific antibodies (BsAbs) therapies	BIOMD0000000788
8	(Dwivedi et al., 2014)	2016 hrs	IBD (Crohn's disease)	BIOMD0000000537
9	(Sluka et al., 2016)	40 mins	Acute Liver Failure	BIOMD0000000624
10	(de Pillis et al., 2008)	5 days	Generic solid tumors (tumor immune interactions)	BIOMD0000000913
11	(León-Triana et al., 2021b)	60 days	Cancer (B-cell acute lymphoblastic leukemia)	BIOMD0000001011
12	(León-Triana et al., 2021b)	1000 days	Cancer (B-cell acute lymphoblastic leukemia)	BIOMD0000001012
13	(León-Triana et al., 2021a)	360 days	Solid Cancer (Glioblastoma)	BIOMD0000001014

Table 6: Rationale for choosing the species as a regimen.

#	Reference	Drug variables	Rationale	Regimen

1	Proprietary Model	Blood.AntiIL23_Central;
		Blood.AntiTNFa_SC;
		Blood.AntiIL1A_central;
		Blood.NBforTNFaOX40L_SC
			Drug variable identified from SimBiology project file	Untreated; 200 mg at days 0, 28, 56 and 100 mg on days 112 and every 56 days thereafter ; 200 mg at days 0 and 100 mg at days 14 and 100 mg at days 42, 70 ; 500 mg at days 0, 14, 28, 42, 56, 70 and 84 ; 150 mg at every 14 days

2	(Gadkar et al., 2022)	PI3Kinh_gi ; PI3Kinh_circ	Drug variable identified from SimBiology project file	Untreated; Umbralisib (800mg once daily); Alpelisib (300mg once daily); Taselisib (4mg once daily); Idelalisib (150mg twice daily); Duvelisib (25mg twice daily); Pictilisib (340mg once daily); Copanlisib (60mg once weekly)
3	(Faratian et al., 2009)	Per	Drug variable identified by looking at which variable corresponding to drug (based on annotation - Pertuzumab) and also 'per' is changed here to plot. Directly given in biomodels	Placebo; Initial concentration of Per is 300000
4	(Benson et al., 2014)	PFM_gut	PFM_gut was connected to dose in MD assignment global quantity	Initial Concentrations (PF-04457845-dose 0.1mg; PF-04457845-dose 0.3mg; PF-04457845-dose 10mg; PF-04457845-dose 1mg; PF-04457845-dose 20mg; PF-04457845-dose 3mg; PF-04457845-dose 40mg)
5	(Wajima et al., 2009)	A_warf	Drug variable identified from events in the SBML file	4mg everyday
6	(Palmér et al., 2014)	Anakinrasc	Drug variable identified from events in the SBML file	Placebo; 200 mg subcutaneously once daily for 13 weeks; 200 mg subcutaneous intermittent dosing regimen alternating between once daily for 13 weeks followed by 13 weeks off treatment
7	(Schropp et al., 2019)	C_free	Drug variable identified by looking at which variable corresponding to drug	Initial concentration of 50mg; Initial concentration of 250mg; 50mg every 2 weeks; 250mg every 2 weeks
8	(Dwivedi et al., 2014)	Ab{serum}	Drug variable identified from events in the SBML file	Placebo; 560mg for every 2 weeks; 560mg for every 4 weeks
9	(Sluka et al., 2016)	APAP	Drug variable identified from Biomodels description	Initial concentration of 0.1
10	(de Pillis et al., 2008)	I	Drug variable identified from Biomodels description	Initial concentration of 2000
11	(León-Triana et al., 2021b)	CAR_T-cells	Drug variable identified from Biomodels description	Initial concentration of 10^7
12	(León-Triana et al., 2021b)	CAR_T-cells	Drug variable identified from Biomodels description	Initial concentration of 10^7
13	(León-Triana et al., 2021a)	CAR_T-cells_off-tumour	Drug variable identified from Biomodels description	Initial concentration of 2 * 10^8

C Supplementary Equations

Z-score transformation was used to transform species onto a similar scape as follows,

$$z_{s,t}^{(p)} = \frac{x_{s,t}^{(p)} - \mu_s}{\sigma_s} \tag{1}$$

where $x_{s,t}^{(p)}$ denotes the value of species s at time t for virtual patient p, while μ_s and σ_s indicate the global mean and standard deviation, respectively.

Symmetric Mean Absolute Percentage Error (SMAPE) is defined as:

$$\text{SMAPE} = \frac{2}{n}\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|} \tag{2}$$

where y and $\hat{y}$ denote the ground-truth and forecast signals, respectively.

Mean Absolute Error (MAE) is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3}$$

where y and $\hat{y}$ denote the ground-truth and forecast signals, respectively.

Mean Squared Error (MSE) is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4}$$

where y and $\hat{y}$ denote the ground-truth and forecast signals, respectively.

Root Mean Squared Error (RMSE) is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{5}$$

where y and $\hat{y}$ denote the ground-truth and forecast signals, respectively.

Prediction Interval Coverage Probability (PICP) is defined as:

$$\text{PICP} = \frac{1}{n}\sum_{i=1}^{n} \nVdash(\hat{y}_i^{lower} \leq y_i \leq \hat{y}_i^{upper}), \tag{6}$$

where $\nVdash$ returns 1 if condition match, otherwise 0. y, $\hat{y}^{lower}$, and $\hat{y}^{upper}$ denote the ground-truth signal and the lower and upper prediction interval bounds, respectively.

Mean Prediction Interval Width (MPIW) is defined as:

$$\text{MPIW} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i^{upper} - \hat{y}_i^{lower} \tag{7}$$

where $\hat{y}^{lower}$ and $\hat{y}^{upper}$ denote the lower and upper prediction interval bounds, respectively.