Temporal Evaluation of Uncertainty Quantification under Distribution Shift

Emma Svensson, Hannah Rosa Friesacher, Adam Arany, Lewis Mervin, Ola Engkvist

Molecular AI and Discovery Sciences, R&D
AstraZeneca, Gothenburg Sweden
ELLIS Unit Linz, Institute for machine learning
Johannes Kepler University Linz, Austria

2024-09-20

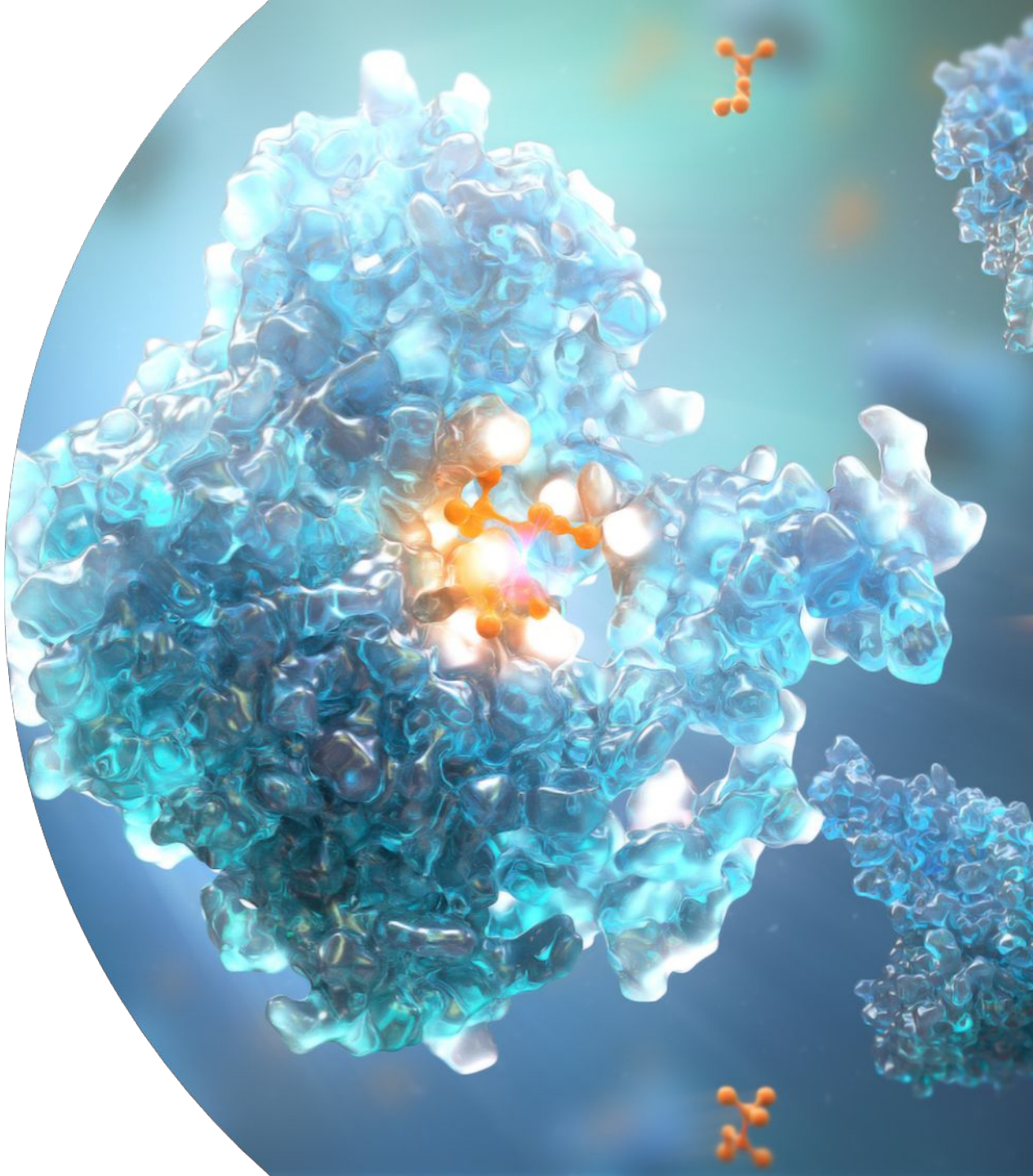


Fig: Biocatalysis enabling efficient and sustainable synthesis of AstraZeneca drug molecules.

Context and Research Questions

High-stakes decisions require UQ in drug discovery (Mervin et al., 2021)

1. *initial screening of chemical space,*
2. *guiding the search, and*
3. *choosing final drug candidate for clinical trial.*

Temporal split most accurately simulates the real drug discovery (Yin et al., 2023).

Low-data problem makes it difficult to accurately evaluate UQ (Hirschfeld et al., 2020).

Censored data is typically available but currently not used in UQ for drug discovery applications (Hüttel et al.,2024).

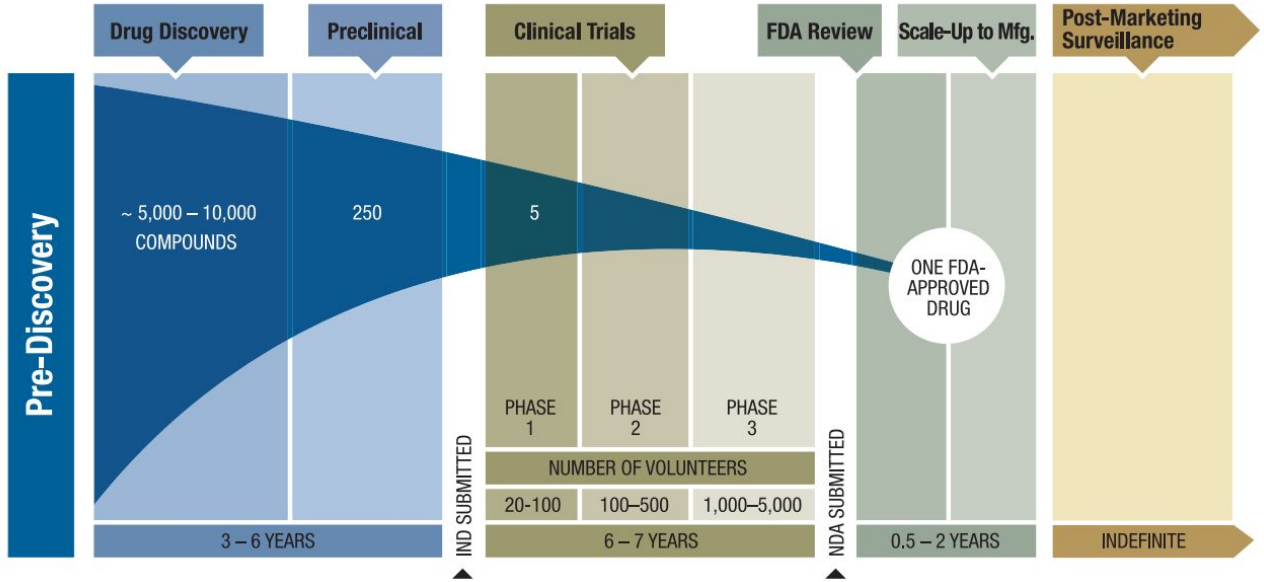


Fig from AACR Cancer Progress Report 2011 Transforming Patient Care Through Innovation.

Hirschfeld, L., et al. "Uncertainty quantification using neural networks for molecular property prediction." *Journal of Chemical Information and Modeling* 60.8 (2020): 3770-3780.
Mervin, L. H., et al. "Uncertainty quantification in drug design." *Drug discovery today* 26.2 (2021): 474-489.
Yin, T., et al. “Evaluating uncertainty-based active learning for accelerating the generalization of molecular property prediction.” *Journal of Cheminformatics* 15, 105 (2023).
Hüttel, F. B., et al. "Bayesian Active Learning for Censored Regression." *arXiv preprint arXiv:2402.11973* (2024).

Temporal Evaluation of Assay-based Data

Assay categories,

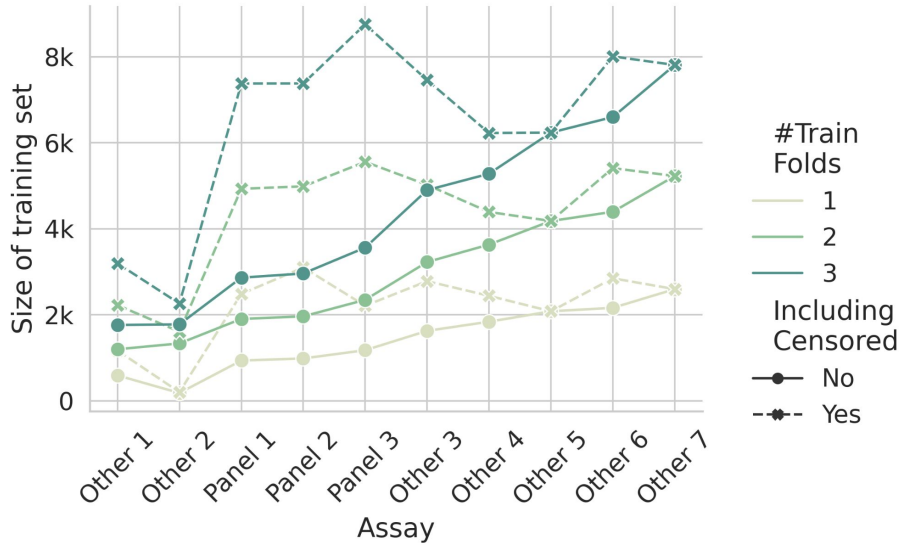
- Panel: e.g. Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADME-T). Cross-project assays for off-target effects.
- Other: project-specific on-target effects.

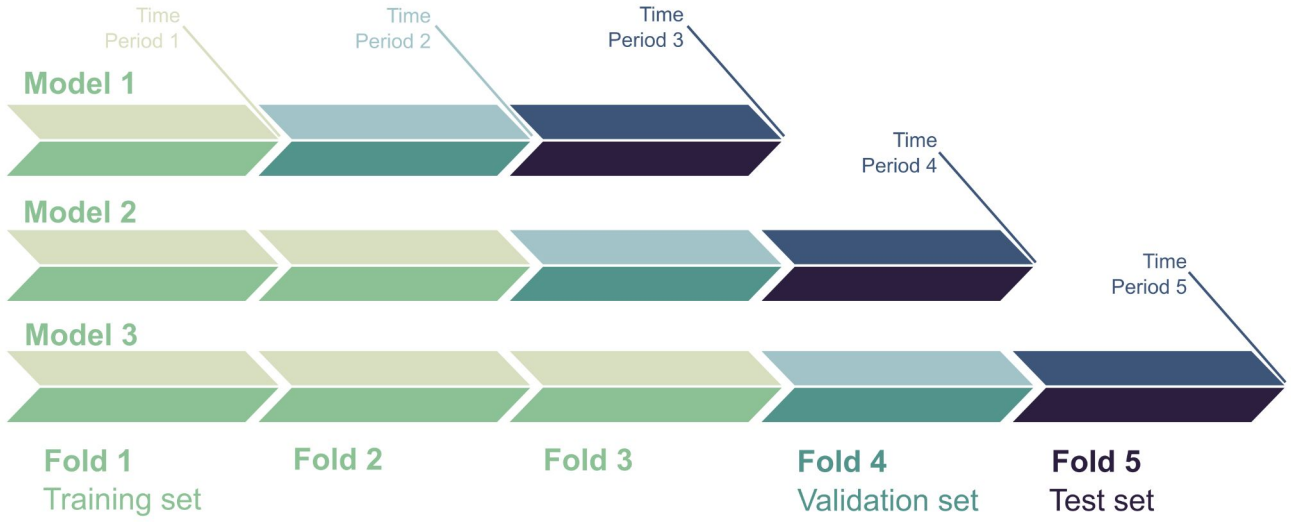
Preprocessing,

- Aggregate duplicated measurements with median
- Transform to log-scale for all end-points
- ECFP with size 1024 and radius 2

Temporal evaluation,

- Split data into five folds based on time
- Train/valid/test on resulting 3 settings



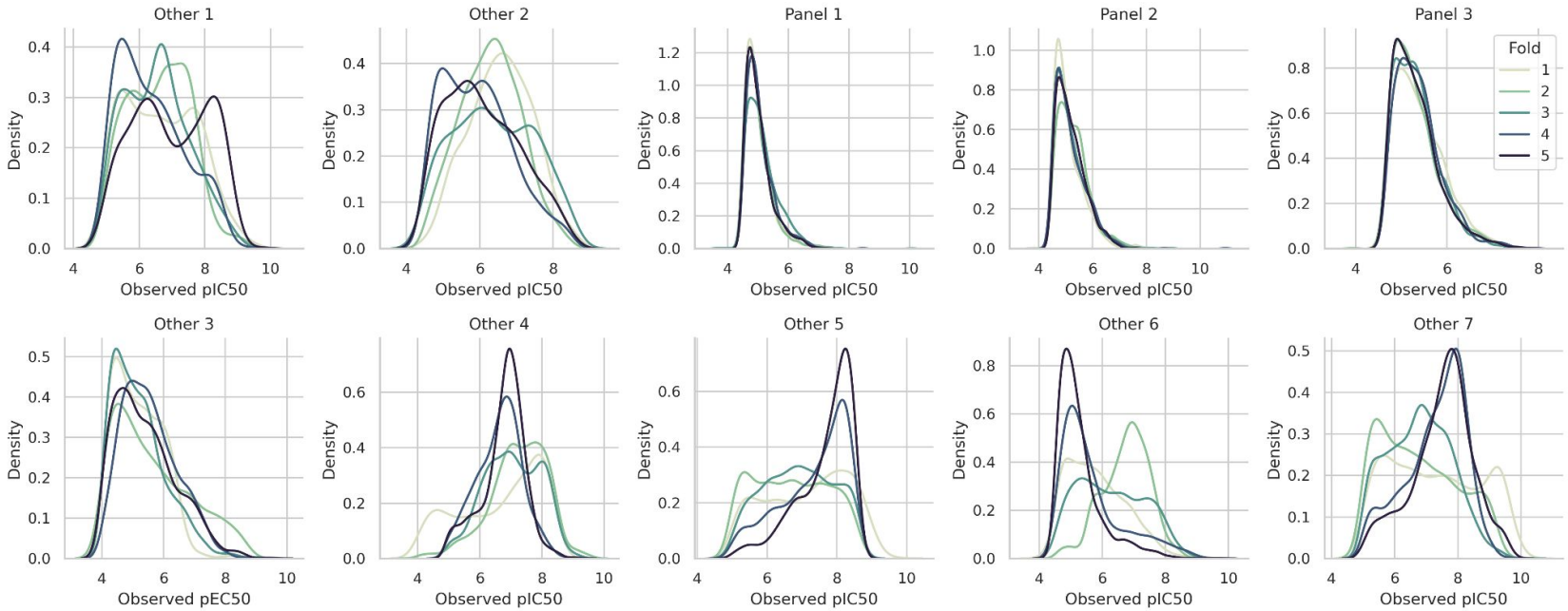


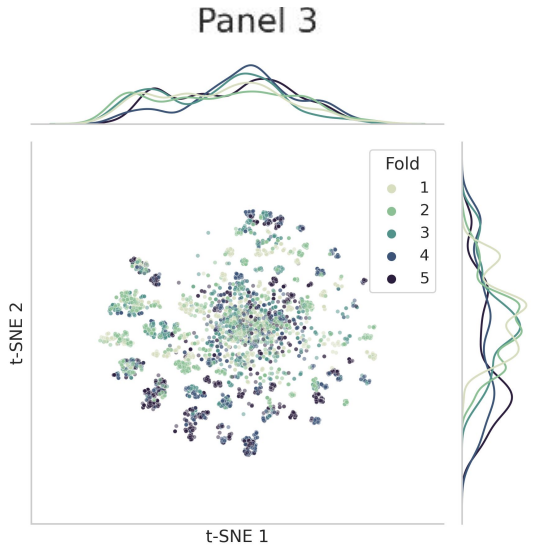
Heyndrickx, W., et al. "MELLODDY: Cross-pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR without Compromising Proprietary Information." J. of Chem. Inf. Model (2023).

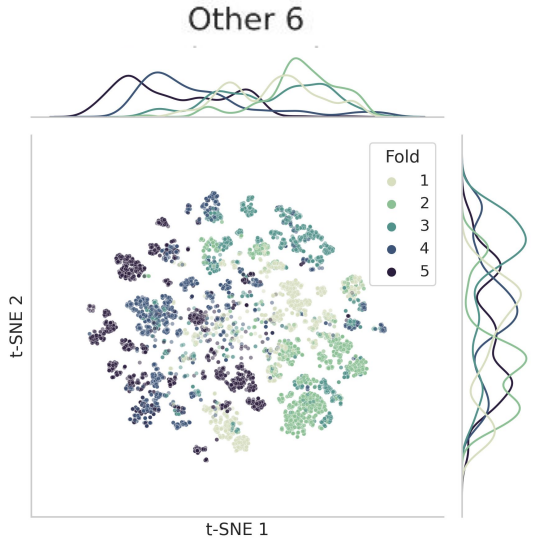
Friesacher, H. R., et al. "Towards Reliable Uncertainty Estimates for Drug Discovery: A Large-scale Temporal Study of Probability Calibration." *ICML 2024 AI for Science Workshop.* (2024)

Temporal Evaluation of Assay-based Data

Data analysis, feature-space and label-space shifts over time…







Modeling UQ in Regression

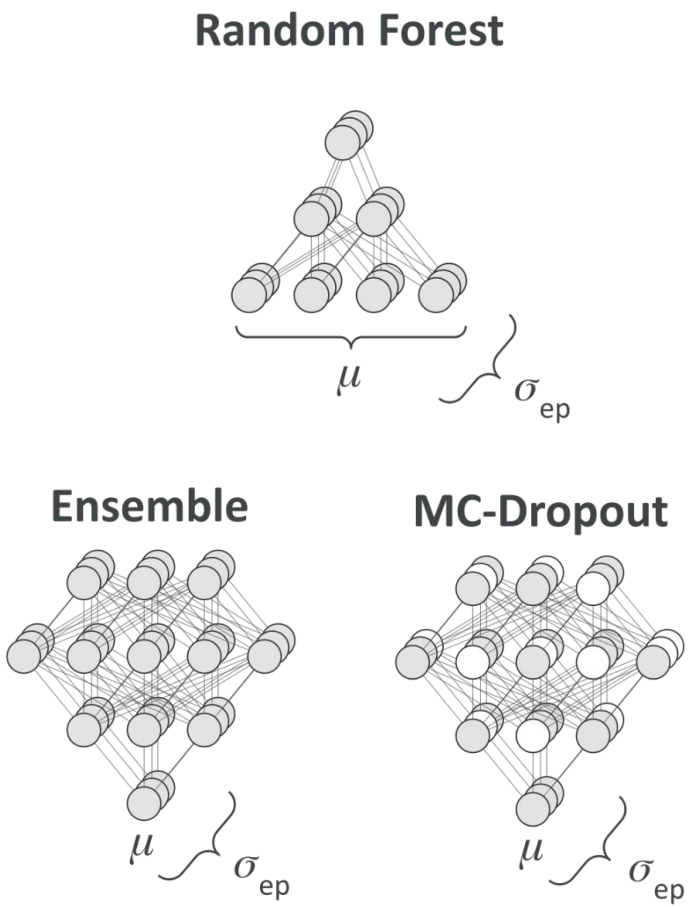
Ensemble-based approaches train multiple individual models independently, e.g. decision trees in Random Forest (Sheridan, 2012) and neural networks (Lakshminarayanan et al., 2017).

Objective, for regression mean squared error (MSE) is used to train each base model. For censored labels we use (Arany et al., 2022),

$$\mathcal{L}^{\text{MSE}} = \frac{1}{N}\sum_{n=1}^{N}\varepsilon_n^2 \qquad \varepsilon_n = \begin{cases} \min\left(z_n - \mu_t(\mathbf{x}_n), 0\right), & \text{if } m_n = -1, \\ y_n - \mu(\mathbf{x}_n), & \text{if } m_n = 0, \\ \max\left(z_n - \mu(\mathbf{x}_n), 0\right), & \text{if } m_n = 1, \end{cases}$$

Result,

- Final prediction is the average over all individual predictions, $\mathbb{E}[\mu]$
- Predicted epistemic uncertainty is the variance between the predictions, $\sigma^2_{\text{ep}} = \text{Var}[\mu]$



Sheridan, R. P. "Three useful dimensions for domain applicability in QSAR models using random forest." *Journal of Chemical Information and Modeling* 52.3 (2012): 814-823.
Gal, Y., and Ghahramani, Z. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." *ICML.* PMLR, 2016.
Lakshminarayanan, B., et al. "Simple and scalable predictive uncertainty estimation using deep ensembles." *NeurIPS* 30 (2017).
Arany, A., et al. "SparseChem: Fast and accurate machine learning model for small molecules." *arXiv preprint arXiv:2203.04676* (2022).

Experimental Setup

Training details,

Optimized hyperparameters for Random Forest, base neural network for Ensemble, MC-Dropout.

Make 10 repeated experiments for all models.

Evaluation metrics,

- Predictive accuracy in terms of MSE
- Calibration of uncertainty with confidence-based calibration curves (Hubschneider et al., 2019)
- Intertwined, overall performance in terms of NLL and ENCE (Levi et al., 2022)

Only key results presented here…

Ablation study,

Compare each model trained with and without censored labels in addition to the observed values.

Model comparison,

Compare the resulting models with each other and the Random Forest baseline.

Case study,

Deeper look at the predicted epistemic uncertainty by the best performing model on an assay with large distribution shifts.

Hubschneider, C., Hutmacher, R., and Zöllner, J. M. "Calibrating uncertainty models for steering angle estimation." *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019.
Levi, D., et al. "Evaluating and calibrating uncertainty prediction in regression tasks." *Sensors* 22.15 (2022): 5540.

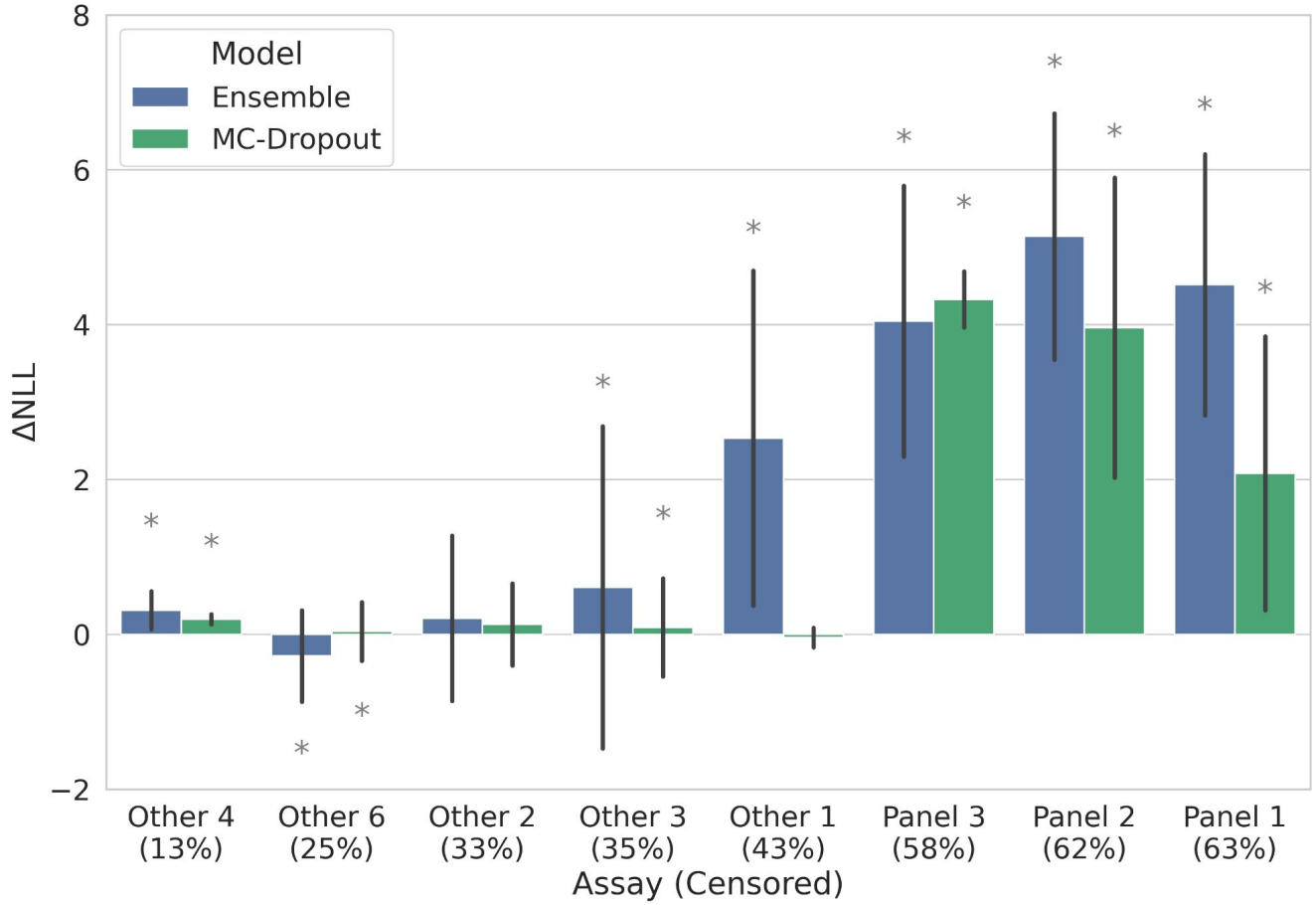
Ablation Study

NLL is adjusted to censored labels using the Tobit model (Tobin, 1958),

$$\mathrm{NLL} = -\frac{1}{N}\sum_{n=1}^{N}(1-|m_n|)\log\varphi(y_n|\mathbf{x}_n,\theta) + |m_n|\log\begin{cases}\Phi(z_n|\mathbf{x}_n,\theta), & \text{if } m_n=-1,\\ 1-\Phi(z_n|\mathbf{x}_n,\theta), & \text{if } m_n=1.\end{cases}$$

Compare, $\Delta\mathrm{NLL} = \mathrm{NLL}_{\mathrm{Observed}} - \mathrm{NLL}_{\mathrm{Censored}}$

Significance is marked with star above/below if censored/observed model is significantly better for majority of settings.

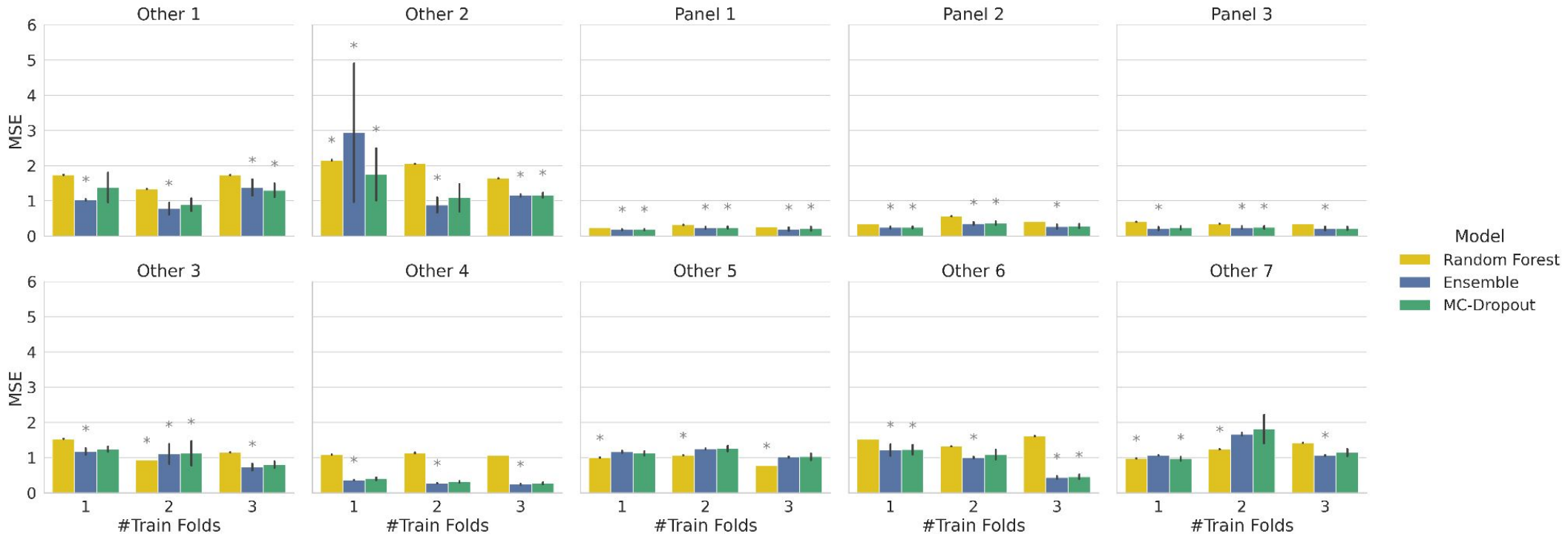


Tobin, J. "Estimation of relationships for limited dependent variables." *Econometrica: Journal of the Econometric Society* (1958): 24-36.

Model Comparison

Overall predictive performance,

Higher accuracy for all models on Panel (ADME-T) assays without distributions shifts.

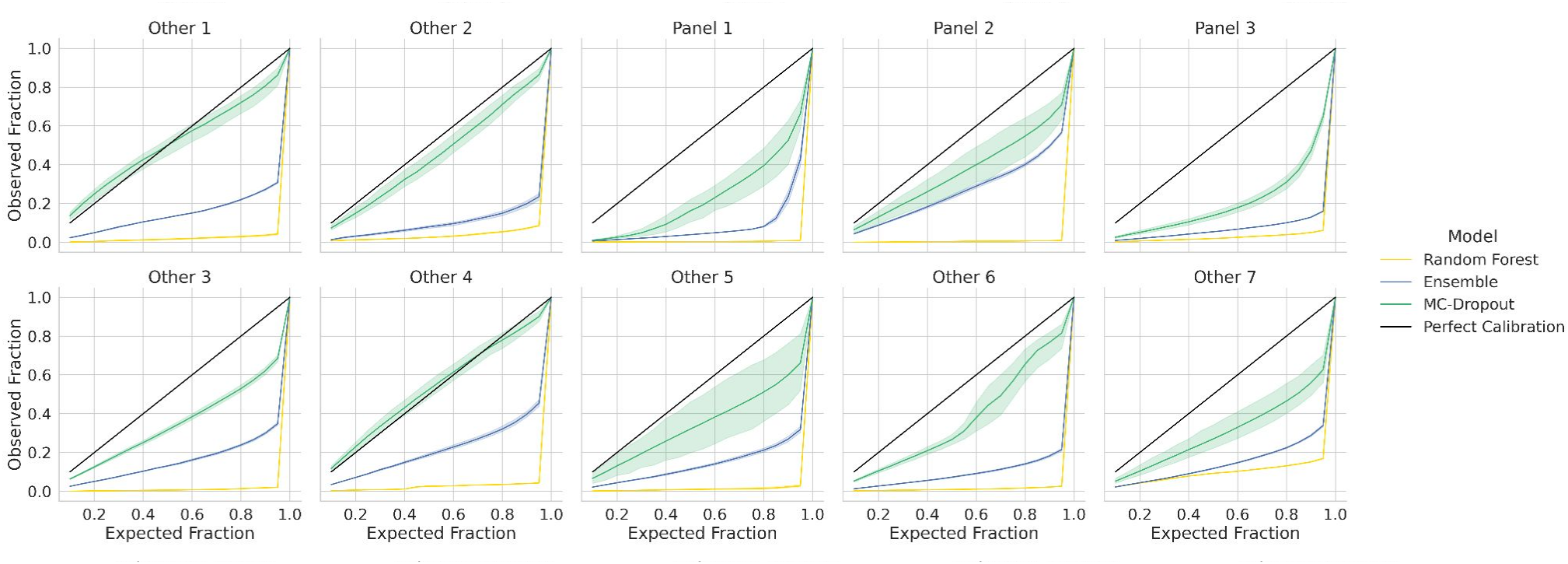


Model Comparison

Confidence-based calibration curves,

1. Convert predicted uncertainty to Confidence Interval (CI), e.g. 1.96σ for a 95% CI.

2. For every predicted z% CI (expected) check fraction of predictions that lie within the corresponding CI (observed).



Conclusions

Censored labels,
Enhance the robustness and reliability of the models, especially when >33% of the available labels are censored.

Best methods,
The highest predictive accuracy varies between assays, but the computationally efficient, Bayesian MC-Dropout model produces consistently better calibrated uncertainty estimates.

Temporal evaluation,
Results from the model comparison are typically robust through time for Panel (ADME-T) assays, where no shifts occur due to the diverse nature of the cross-project assays.
For target-based assays, it can change drastically, requiring re-evaluated comparison from time to time.

In our extended work, we have added Bayesian and Gaussian models as well as more ADME-T assays (Svensson et al., 2024)

Svensson, E., et al. "Enhancing Uncertainty Quantification in Drug Discovery with Censored Regression Labels." arXiv preprint arXiv:2409.04313 (2024).

Thank you!
Questions?

Acknowledgments

Rosa Friesacher
Susanne Winiwarter
Lewis Mervin
Ola Engqvist

Adam Arany

Günter Klambauer
Sepp Hochreiter

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant No 956832.

References

Tobin, J. "Estimation of relationships for limited dependent variables." Econometrica: Journal of the Econometric Society (1958): 24-36.

Sheridan, R. P. "Three useful dimensions for domain applicability in QSAR models using random forest." *Journal of Chemical Information and Modeling* 52.3 (2012): 814-823.

Gal, Y., and Ghahramani, Z. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." *International conference on machine learning*. PMLR, 2016.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. "Simple and scalable predictive uncertainty estimation using deep ensembles." *Advances in neural information processing systems* 30 (2017).

Hubschneider, C., Hutmacher, R., and Zöllner, J. M. "Calibrating uncertainty models for steering angle estimation." *2019 IEEE intelligent transportation systems conference (ITSC)*. IEEE, 2019.

Hirschfeld, L., et al. "Uncertainty quantification using neural networks for molecular property prediction." *Journal of Chemical Information and Modeling* 60.8 (2020): 3770-3780.

Mervin, L. H., et al. "Uncertainty quantification in drug design." *Drug Discovery Today* 26.2 (2021): 474-489.

Arany, A., et al. "SparseChem: Fast and accurate machine learning model for small molecules." *arXiv preprint arXiv:2203.04676* (2022).

Levi, D., et al. "Evaluating and calibrating uncertainty prediction in regression tasks." *Sensors* 22.15 (2022): 5540.

Heyndrickx, W., et al. "MELLODDY: Cross-pharma Federated Learning at Unprecedented Scale Unlocks Benefits in QSAR without Compromising Proprietary Information." J. of Chem. Inf. Model (2023).

Yin, T., Panapitiya, G., Coda, E.D., et al. “Evaluating uncertainty-based active learning for accelerating the generalization of molecular property prediction.” *Journal of Cheminformatics* 15, 105 (2023).

Friesacher, H. R., et al. "Towards Reliable Uncertainty Estimates for Drug Discovery: A Large-scale Temporal Study of Probability Calibration." *ICML 2024 AI for Science Workshop.* (2024).

Hüttel, F. B., et al. "Bayesian Active Learning for Censored Regression." *arXiv preprint arXiv:2402.11973* (2024).

Svensson, E., et al. "Enhancing Uncertainty Quantification in Drug Discovery with Censored Regression Labels." arXiv preprint arXiv:2409.04313 (2024).

Read our extended paper: