

Designing formulations of bio-based, multi-component epoxy resin systems via machine learning

Rodrigo Q. Albuquerque,¹ Florian Rothenhäusler,¹ and Holger Ruckdäschel^{1,2*}

¹ Department of Polymer Engineering, University of Bayreuth, Universitätsstraße 30, Bayreuth, 95447, Germany.

² Neue Materialien Bayreuth GmbH, Gottlieb-Keim-Straße 60, Bayreuth, 95448, Germany.

* E-mail(s): holger.ruckdaeschel@uni-bayreuth.de

Supporting Information (SI)

- Detailed evaluation of the performance of all ML models discussed in the main text. The models' performances are also shown in Fig. 6 of the main text.
- Case study of the hyperparameter optimization of the small, 35-sample dataset using the KRR model as example.

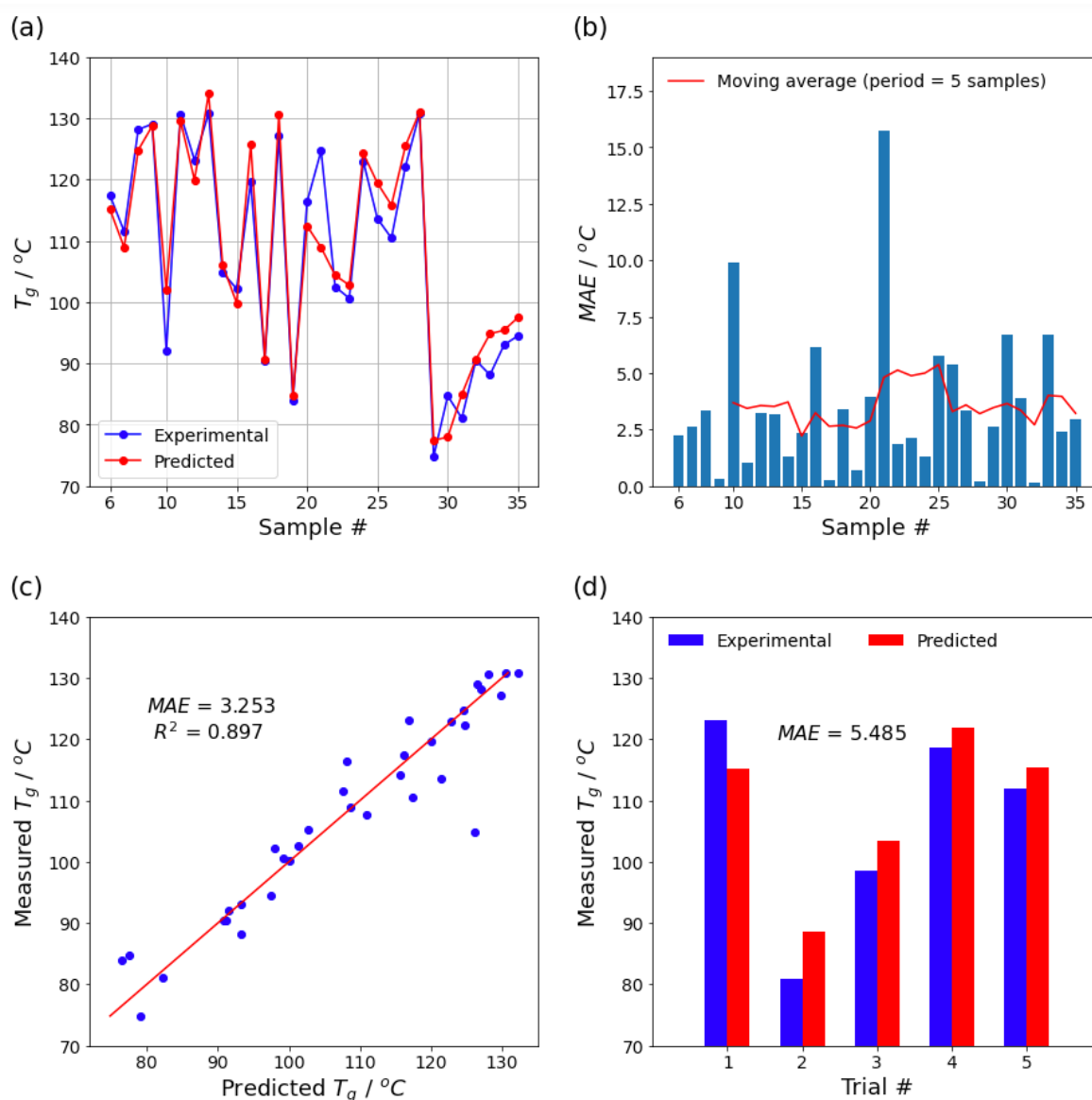


Figure S1. a) Comparison between predicted and experimental T_g . The prediction for sample i was performed by training a fresh model using samples 1 to $i-1$. b) Error for the predictions shown in a) (bars) and their moving average (line). c) 5-fold cross-validation evaluation of the performance of the model including all 35 samples. d) Comparison between the predicted and experimental T_g for the experimental validation set of randomly selected formulations. The predictions shown in all subplots were performed using the KRR model with default, non-optimized hyperparameters ($\alpha = 0.001$, $\gamma = 0.01$, kernel = 'rbf').

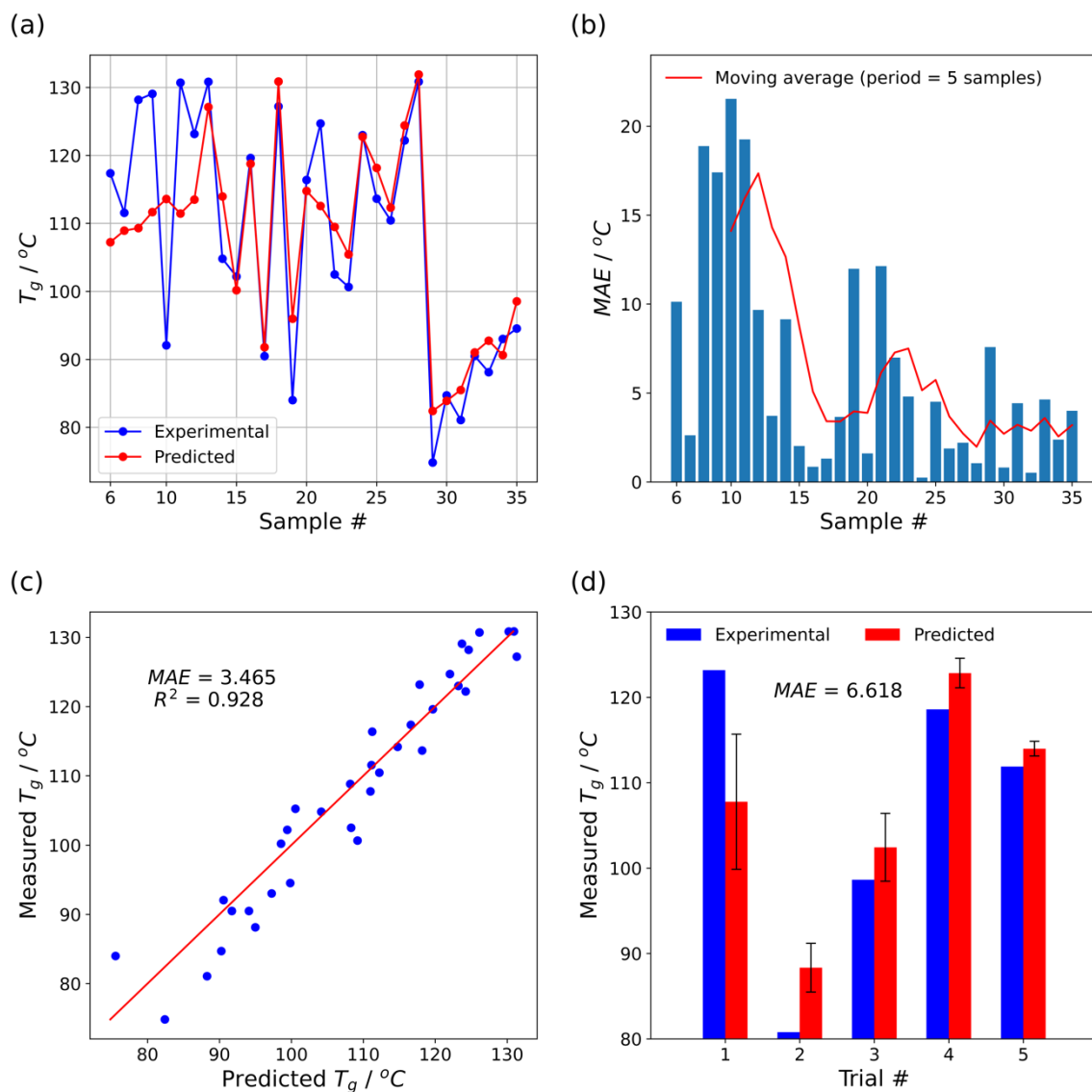


Figure S2. a) Comparison between predicted and experimental T_g . The prediction for sample i was performed by training a fresh model using samples 1 to $i-1$. b) Error for the predictions shown in a) (bars) and their moving average (line). c) 5-fold cross-validation evaluation of the performance of the model including all 35 samples. d) Comparison between the predicted and experimental T_g for the experimental validation set of randomly selected formulations. The predictions shown in all subplots were performed using the non-parametric GPR model.

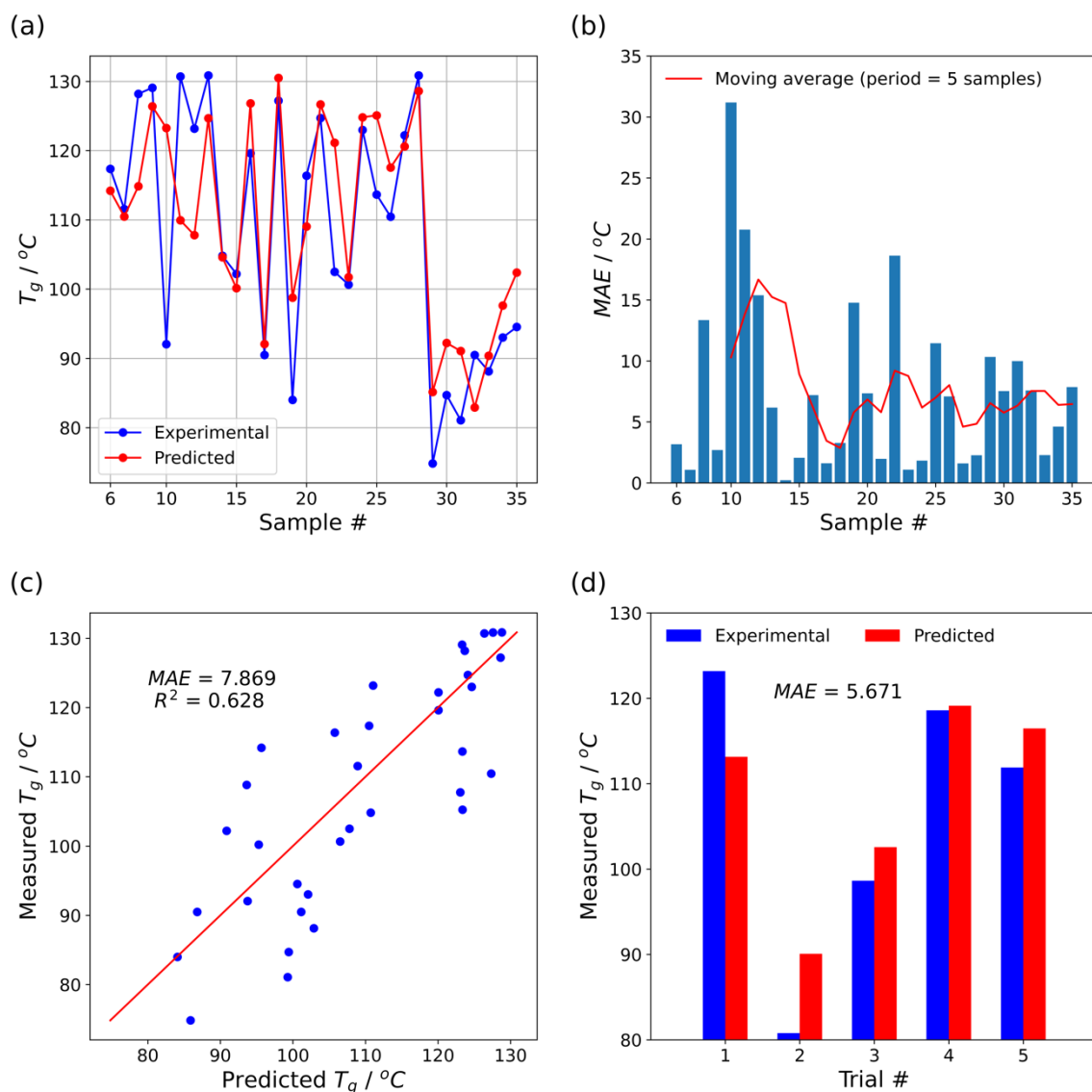


Figure S3. a) Comparison between predicted and experimental T_g . The prediction for sample i was performed by training a fresh model using samples 1 to $i-1$. b) Error for the predictions shown in a) (bars) and their moving average (line). c) 5-fold cross-validation evaluation of the performance of the model including all 35 samples. d) Comparison between the predicted and experimental T_g for the experimental validation set of randomly selected formulations. The predictions shown in all subplots were performed using a GBR model with default, non-optimized hyperparameters (loss = 'ls', learning_rate = 0.1).

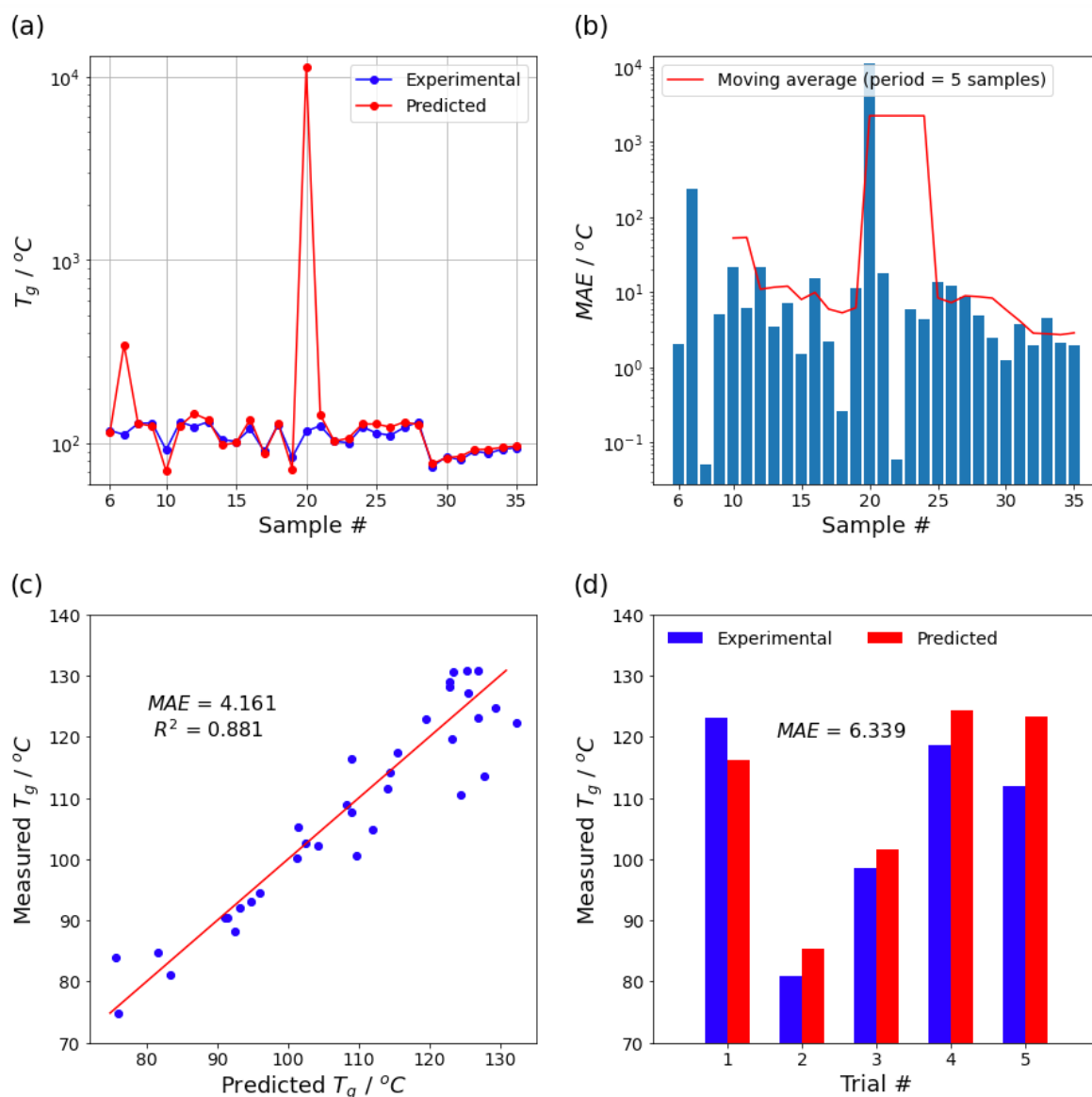


Figure S4. a) Comparison between predicted and experimental T_g . The prediction for sample i was performed by training a fresh model using samples 1 to $i-1$. b) Error for the predictions shown in a) (bars) and their moving average (line). c) 5-fold cross-validation evaluation of the performance of the model including all 35 samples. d) Comparison between the predicted and experimental T_g for the experimental validation set of randomly selected formulations. The predictions shown in all subplots were performed using the LS model.

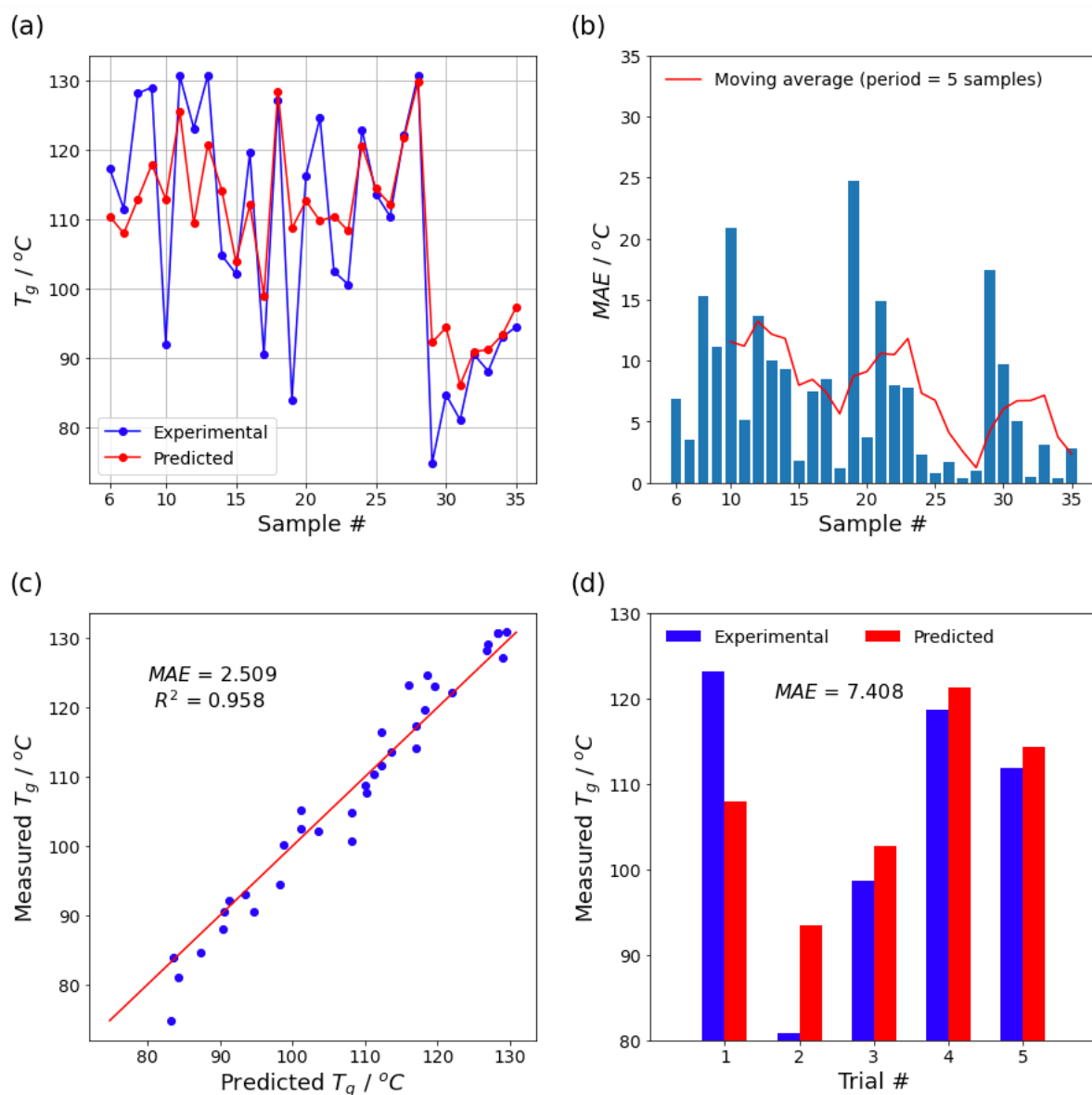


Figure S5. a) Comparison between predicted and experimental T_g . The prediction for sample i was performed by training a fresh model using samples 1 to $i-1$. b) Error for the predictions shown in a) (bars) and their moving average (line). c) 5-fold cross-validation evaluation of the performance of the model including all 35 samples. d) Comparison between the predicted and experimental T_g for the experimental validation set of randomly selected formulations. The predictions shown in all subplots were performed using the SVR model with default, non-optimized hyperparameters (kernel = 'rbf', $C = 1$, epsilon = 0.1).

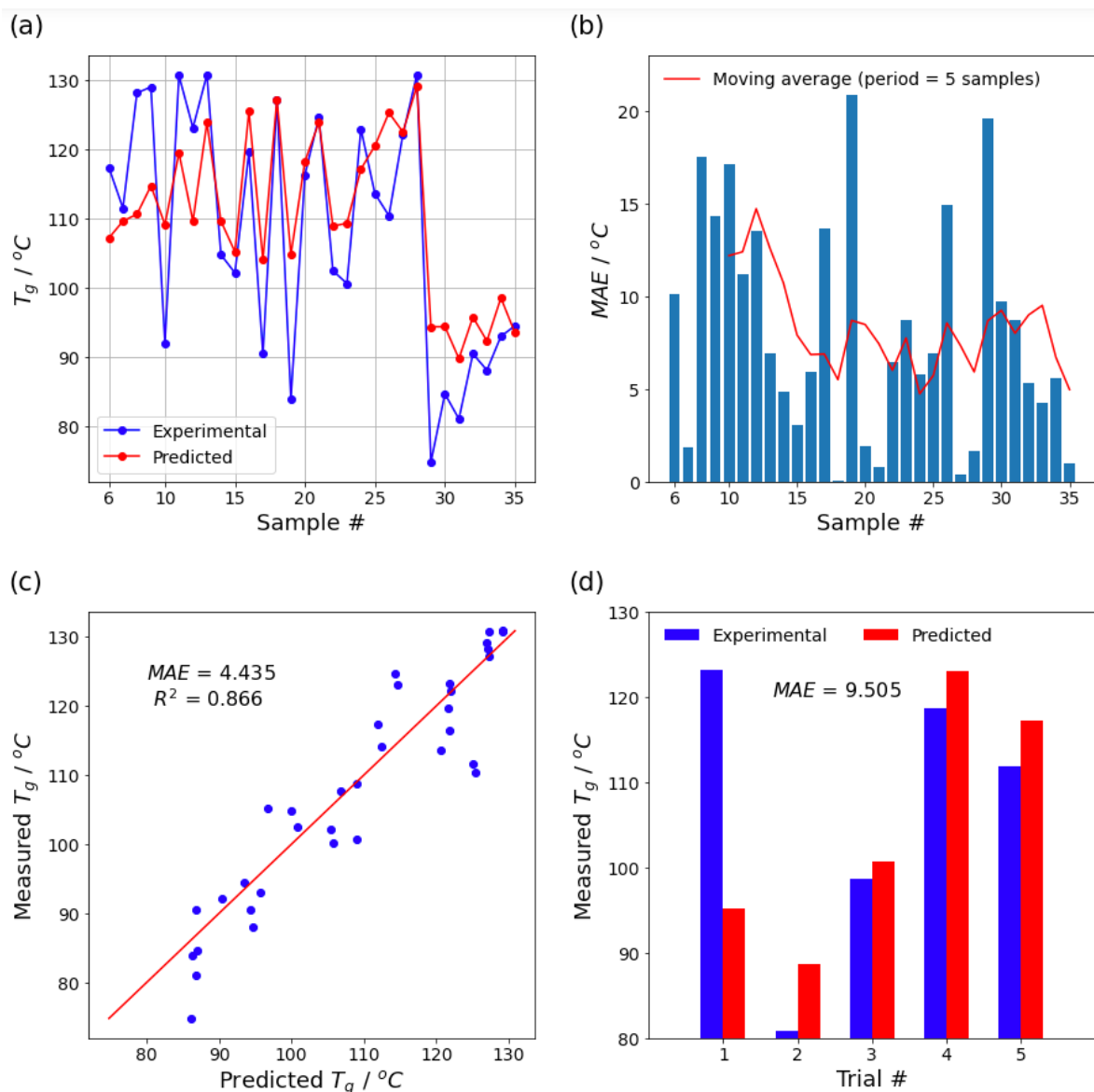


Figure S6. a) Comparison between predicted and experimental T_g . The prediction for sample i was performed by training a fresh model using samples 1 to $i-1$. b) Error for the predictions shown in a) (bars) and their moving average (line). c) 5-fold cross-validation evaluation of the performance of the model including all 35 samples. d) Comparison between the predicted and experimental T_g for the experimental validation set of randomly selected formulations. The predictions shown in all subplots were performed using the KNN model with default, non-optimized hyperparameters ($k_{\text{nearest_neighbors}} = 1$).

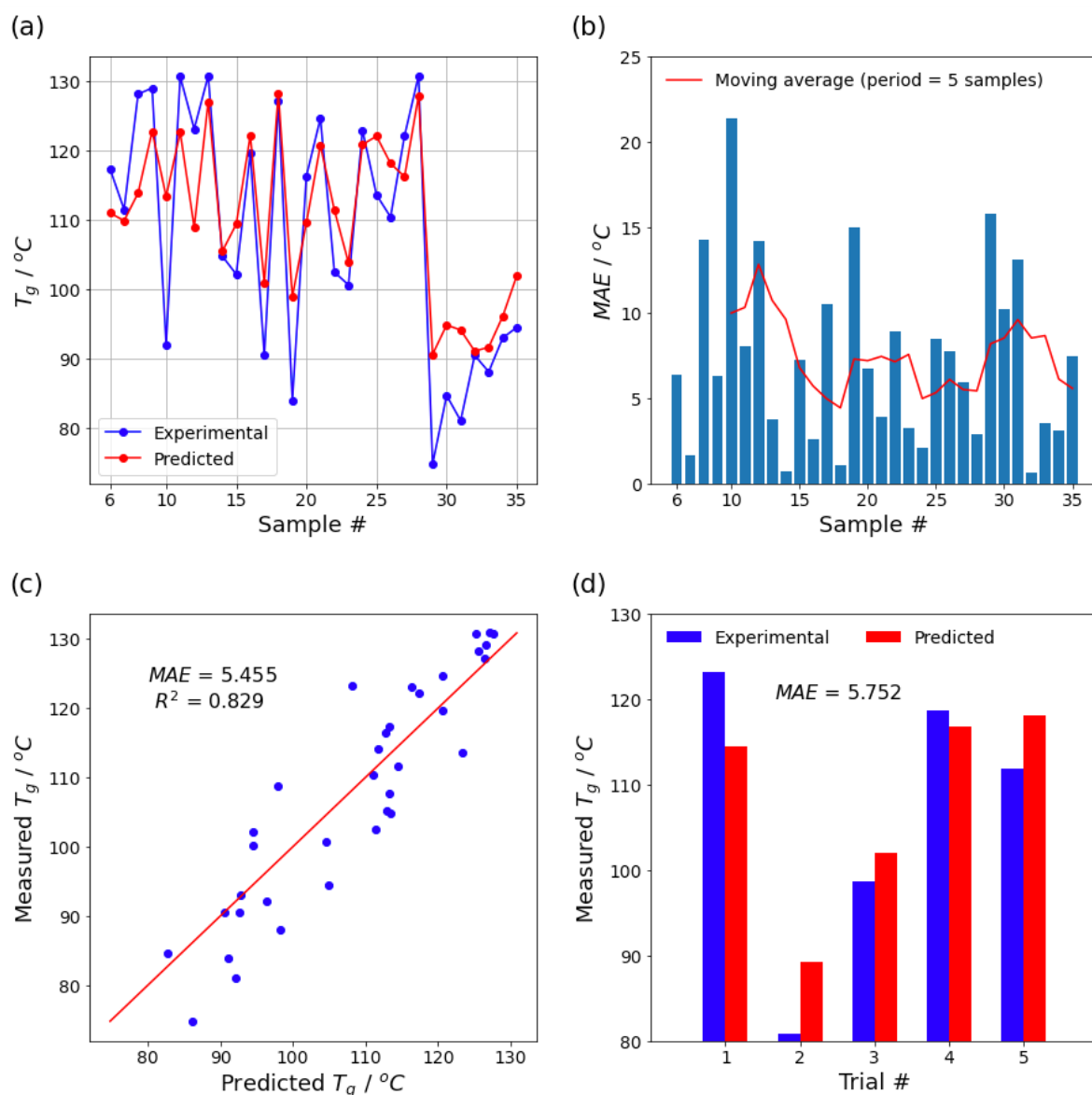


Figure S7. a) Comparison between predicted and experimental T_g . The prediction for sample i was performed by training a fresh model using samples 1 to $i-1$. b) Error for the predictions shown in a) (bars) and their moving average (line). c) 5-fold cross-validation evaluation of the performance of the model including all 35 samples. d) Comparison between the predicted and experimental T_g for the experimental validation set of randomly selected formulations. The predictions shown in all subplots were performed using the RF model with default, non-optimized hyperparameters ($n_trees = 100$).

Hyperparameter optimization for the KRR model

For each row of the table below, the 35 samples were split into training + validation sets (22 samples) and test set (13 samples) using the random seeds shown. The KRR hyperparameters were optimized by training different models with small training sets and minimizing the average prediction error on the different validation sets using a 4-fold CV approach. The final optimized ***alpha*** and ***gamma*** hyperparameters (shown in the table below, kernel = 'rbf') were used to train a KRR model with all 22 samples of the training + validation sets, whose final performance (***MAE*** and ***R*²**) was evaluated on the 13-sample test set (test error), as well as on the training + validation sets (training error). Notice the large variance on the ***MAE*** values calculated for the test sets.

Random seed	<i>alpha</i>	<i>gamma</i>	<i>MAE</i> (training)	<i>R</i> ² (training)	<i>MAE</i> (test)	<i>R</i> ² (test)
10	0.021	0.008	1.773	0.980	3.010	0.894
11	0.023	0.020	1.592	0.987	2.711	0.910
12	0.022	0.016	1.305	0.992	2.435	0.957
13	0.026	0.009	1.653	0.984	4.726	0.854
14	0.065	0.017	2.159	0.972	2.018	0.965
15	0.010	0.007	1.263	0.992	5.388	0.560
16	0.014	0.013	1.414	0.984	2.476	0.951
17	0.038	0.021	1.359	0.984	2.104	0.983
18	0.001	0.014	0.310	0.999	4.648	0.883
19	0.025	0.007	2.069	0.979	5.266	0.537
Column averages (+/-std) =			1.490(0.490)	0.985(0.007)	3.348(1.292)	0.849(0.155)