Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Supporting Information for: Enhanced Thermophysical Property Prediction with Uncertainty Quantification using Group Contribution-Gaussian Process Regression

Barnabas P. Agbodekhe,[a] Montana N. Carlozo,[a] Dinis O. Abranches,[b] Kyla D. Jones,[a] Alexander W. Dowling,*[a] and Edward J. Maginn*[a]

S1 Methods and Data

S1.1 Data Collection and Preparation

S1.1.1 Joback and Reid GC Predictions

Figure S1 shows how the JR GC model in particular (and simple GC models in general) works. Each molecular group has a contribution to the total property value of the molecule. The group contributions $T_{b,i}$ for the normal boiling temperature T_b are obtained as parameters from the JR GC model[1].

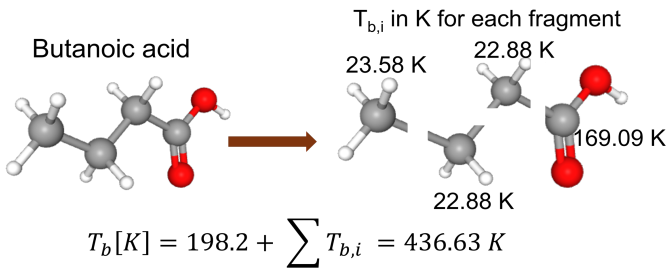


Fig. S1 Illustration of Joback and Reid (JR) GC method

S1.2 Data Pre-processing

S1.2.1 Data Quality

Figures S2 and S3 shows the results of outlier analysis on the entire datasets comparing predictions from the Joback and Reid (JR) GC method with experimental data. Either a power law or linear function (green line) was fitted to the JR GC predictions compared to experimental data, as shown in Figure S2. Two boundary lines (red lines) were then constructed representing two standard deviations of the JR GC predictions from the function fit line for each property. Red points represent suspected outliers. These points were the focus of data quality checks. For all data points flagged by the outlier detection code, we cross-checked the reported values in the CRC Handbook of Chemistry and Physics[2] with values from Yaws' Critical Property Data for Chemical Engineers and Chemists[3] as available in the Knovel database. Comparisons were also made with reported values in the National Institute of Standards and Technology (NIST) webbook as available. For all suspected outlier points, there was agreement between the CRC Handbook of Chemistry and Physics and Yaws' Critical Property Data for Chemical Engineers and Chemists except for three molecules in the ΔH_{vap} dataset. Figure S4 shows the three molecules for which we consider the entries in the CRC Handbook of Chemistry and Physics to be incorrect. We compared the values from the CRC Handbook of Chemistry and Physics with those from Yaws' Critical Property Data for Chemical Engineers and Chemists and data from NIST webbook[4]. For the three molecules shown in Figure S4, the reported values from Yaws' Critical Property Data for Chemical Engineers and Chemists and those from NIST agree with each other while those from the CRC Handbook of Chemistry and Physics are significantly different.

Table S1 shows data from NIST webbook as collated from the original references shown in Table S1 for the three molecules shown in Figure S4. Figure S5 shows the outlier analysis results after the experimental data from the CRC Handbook of Chem-

Table S1 ΔH_{vap} values compiled by National Institute of Standards and Technology (NIST) for molecules with erroneous ΔH_{vap} data from CRC Handbook of Chemistry and Physics. The collated data from NIST and Yaws' critical properties handbook data agree (considering the trend in ΔH_{vap} with temperature) and differ from those of the CRC Handbook of Chemistry and Physics

CAS RN	T /K	ΔH_{vap} /$kJmol^{-1}$	T_b /K	Reference for ΔH_{vap}, T_b
96-48-0	357	51.8	479.20	5,6
	392	49.5		5
	407	48.2		7
590-67-0	355	49.1	428.20	7,8
464-49-3	298.15	55.3	482.15	9,2

[a] Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, IN 46556, USA. E-mail: ed@nd.edu, adowling@nd.edu
[b] CICECO - Aveiro Institute of Materials, Department of Chemistry, University of Aveiro, 3810-193 Aveiro, Portugal.
∗ Corresponding authors

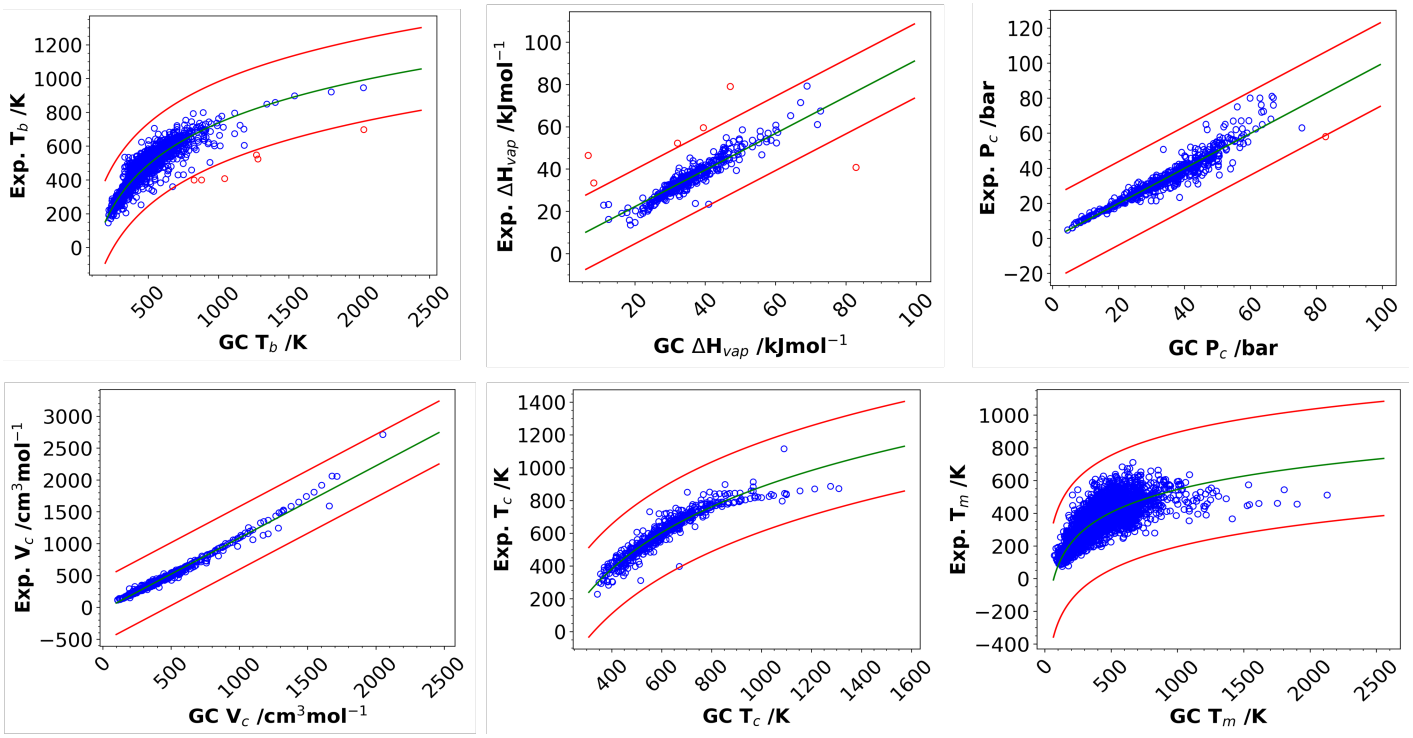


Fig. S2 Outlier analysis using JR GC predictions

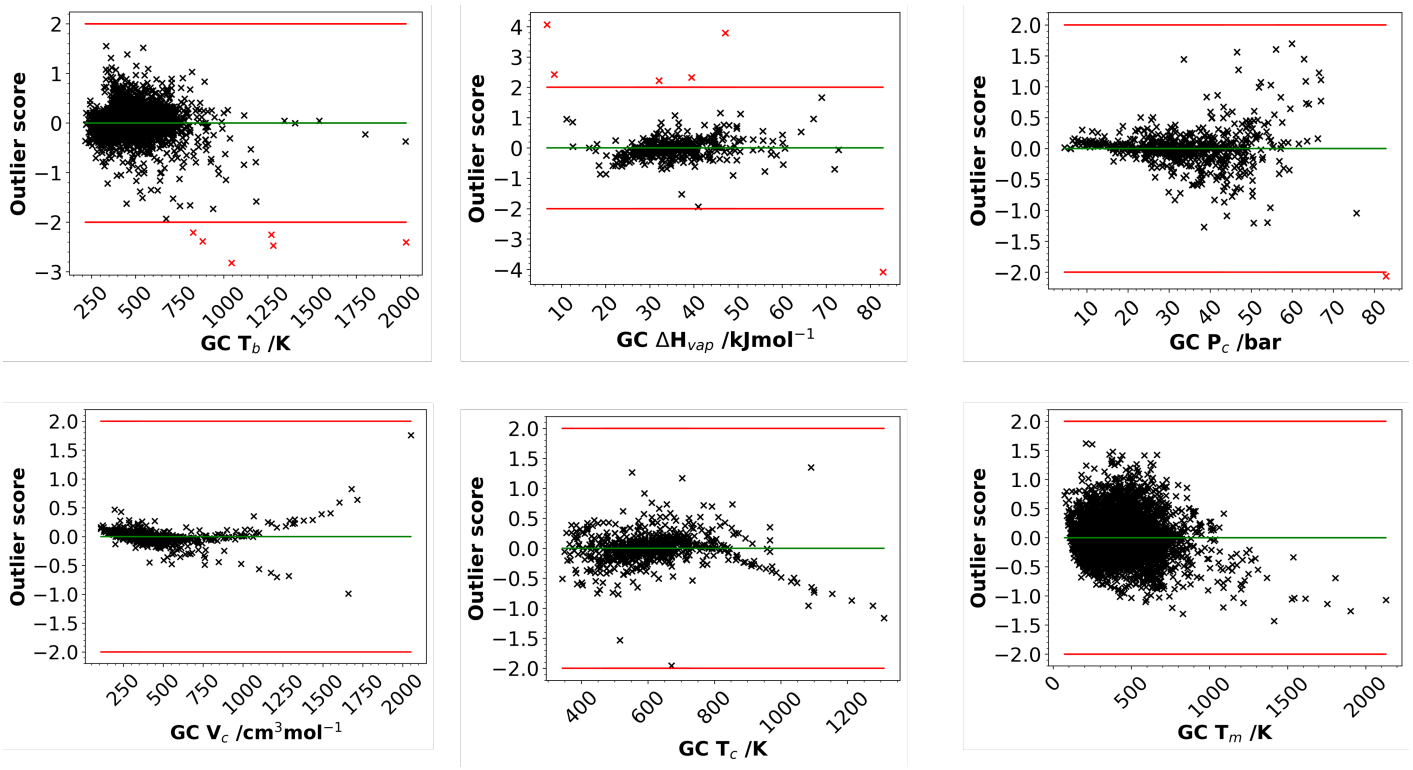


Fig. S3 Outlier scores. Scores scaled by one experimental data standard deviation.

istry and Physics for the three molecules in Figure S4 have been replaced with those from Yaws' Critical Property Data for Chemical Engineers and Chemists. Interestingly, once the experimental data from the CRC Handbook of Chemistry and Physics for the three molecules in Figure S5 were replaced with the correct values from Yaws' Critical Property Data for Chemical Engineers and Chemists, the molecules were no longer flagged as outliers.

There were still however three other molecules for ΔH_{vap} that

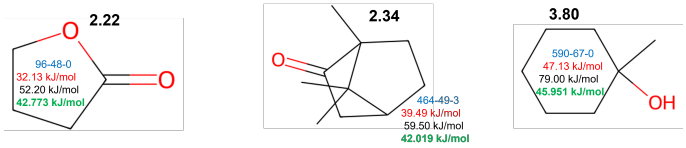


Fig. S4 Molecules with erroneous ΔH_{vap} data from CRC Handbook of Chemistry and Physics. Blue text is CAS number, red is JR GC predictions, black is CRC Handbook data and green is Yaws' critical properties handbook data.

remained flagged as outliers and whose data from the CRC Handbook of Chemistry and Physics were shown to be most likely correct as they agree with data from NIST and Yaws' Critical Property Data for Chemical Engineers and Chemists. These three molecules are shown in Figure S6. It can be observed that these molecules are highly fluorinated or highly nitrated molecules. Detailed discussions have been provided in the main text about why the JR GC model significantly underestimate ΔH_{vap} for highly fluorinated molecules and overestimate ΔH_{vap} for highly nitrated molecules as shown in Figure S5.

S1.2.2 Visualization of Feature vs Label data

S1.3 GP Models

We implemented four separate GP model structures to improve the results of JR GC predictions. Three of the models (Models 1, 3, and 4) require both the GC property predictions $\mathbf{y_{GC}}$ and molecular weights ($\mathbf{M.W}$) and one (Model 2) only requires ($\mathbf{MW}$) as an input. Each model makes different assumptions about the relationship between the input features and experimental property data $\mathbf{y}_{\text{exp}}$. We denote $\mathbf{X} = [\mathbf{y_{GC}}, \mathbf{MW}]$

Model 1 described by equation (S1)

models $\mathbf{y}_{\text{exp}}$ using a mean function $m = 0$. This method is best suited for models where trends in the data are unknown or difficult to elucidate and rely on the kernel function to fully describe trends in the data.

$$\mathbf{y}_{\text{exp}} \sim GP = \mathcal{N}(0, K(\mathbf{X})) \tag{S1}$$

Models 2 and 3 are represented by equations S2 and S3. These models use $m = \mathbf{y_{GC}}$. The difference between equations S2 and S3 is that equation S2 models $\mathbf{y}_{\text{exp}}$ as a function of only $\mathbf{MW}$ while equation S3 models $\mathbf{y}_{\text{exp}}$ as a function of both $\mathbf{y_{GC}}$ and $\mathbf{MW}$.

$$\mathbf{y}_{\text{exp}} \sim GP = \mathcal{N}(\mathbf{y_{GC}}, K(\mathbf{MW})) \tag{S2}$$

$$\mathbf{y}_{\text{exp}} \sim GP = \mathcal{N}(\mathbf{y_{GC}}, K(\mathbf{X})) \tag{S3}$$

In this work, Model 3 is the final model implementation. This method is most suitable for property prediction data where $\mathbf{y_{GC}}$ is at least a somewhat faithful representation of $\mathbf{y}_{\text{exp}}$ and the relationship between MW and $\mathbf{y}_{\text{exp}}$ is unknown.

The fourth GP model form (Model 4) in equation S4 is a generalization of S3 which includes tunable hyperparameters A and b in the mean function such that $m(\mathbf{x}|\mathbf{A}, \mathbf{b}) = \mathbf{Ax} + \mathbf{b}$. This model is well suited for property predictions where the correlation of MW and $\mathbf{y}_{\text{exp}}$ is linear or when $\mathbf{y_{GC}}$ is a poor estimator of $\mathbf{y}_{\text{exp}}$.

$$\mathbf{y}_{\text{exp}} \sim GP = \mathcal{N}(\beta_0 \mathbf{M.W} + \beta_1 \mathbf{y_{GC}} + \beta_2, K(\mathbf{X})) \tag{S4}$$

S2 Results and Discussion

S2.1 Final Model Parameters

Table S2 provide the final hyperparameter values for the final model implementation in this work. The high value of the shape parameter α for P_c indicates that there is a very smooth function mapping features to target for P_c.

Table S2 Final values of the optimized GP hyperparameters for each property

π	l	α	RQ σ	White σ	LML
T_b	4.6367	0.0248	25.0632	0.0915	-913.8
H_{vap}	3.0504	0.1422	9.5645	0.0935	-148.4
P_c	1.9268	4998.8993	0.7850	0.0606	-36.7
V_c	0.6167	0.0194	0.4653	0.0023	658.1
T_c	1.7261	0.1038	2.2611	0.0479	-24.7
T_m	6.3135	0.0216	27.3191	0.2564	-3417.5

For such very smooth functions, the RBF kernel will likely outperform most other kernel architectures while the Matern kernel with $\nu = 1/2$ (suitable for non-smooth functions) will have the worst performance as shown in Table S5. Lower values of α correspond to lower degree of smoothness in general and the Matern kernels can be expected to outperform the RBF kernels for such properties. This is in agreement with the results in Table S5. The RQ kernel represents a general-purpose option that provides competitive performance irrespective of the smoothness of the function that maps input features to targets. Furthermore the additional α parameter of the RQ kernel is physically interpretable as a measure of the degree of smoothness for the function which maps features to targets.

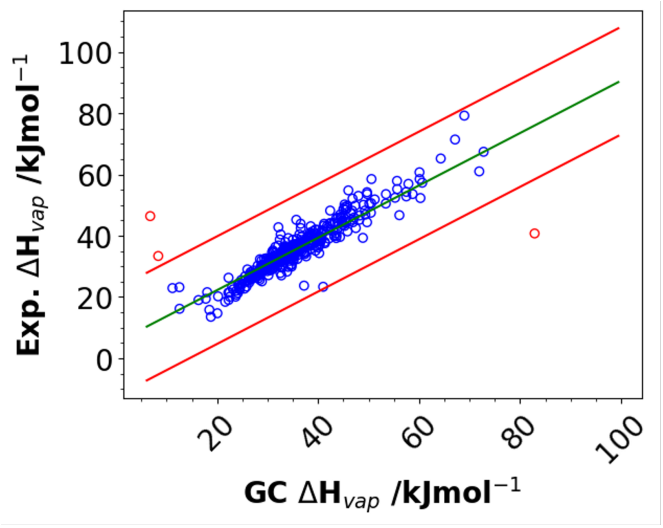
S2.2 Effect of White Kernel Variance settings on T_m Results

Table S3 shows the results for different settings of the white kernel variance parameter on the results for T_m

Table S3 Effect of different settings of the white kernel variance on the model training metrics for T_m. MAPE = Mean Absolute Percentage Error. MAE = Mean Absolute Error. a = Start at $1e^{-5}$ and optimize, b and c = Fix at value corresponding to 95 % confidence interval of experimental uncertainty of 1 K and 10 K respectively, d = Final Implementation used in this work (Start at 1.0 and optimize)

	Train			Test		
Setting	R^2	MAPE	MAE	R^2	MAPE	MAE
a	0.75	12.55	39.66	0.73	12.36	40.57
b	0.98	1.99	6.44	0.6	13.94	46.25
c	0.98	2.04	6.60	0.6	13.94	46.25
d	0.75	12.55	39.66	0.73	12.36	40.57

For options a and d which entail starting from an initial value of $1e^{-5}$ and 1 respectively for the white kernel variance parameter and then optimizing, we see that the final optimized results are similar. However, when the white noise variance parameter is



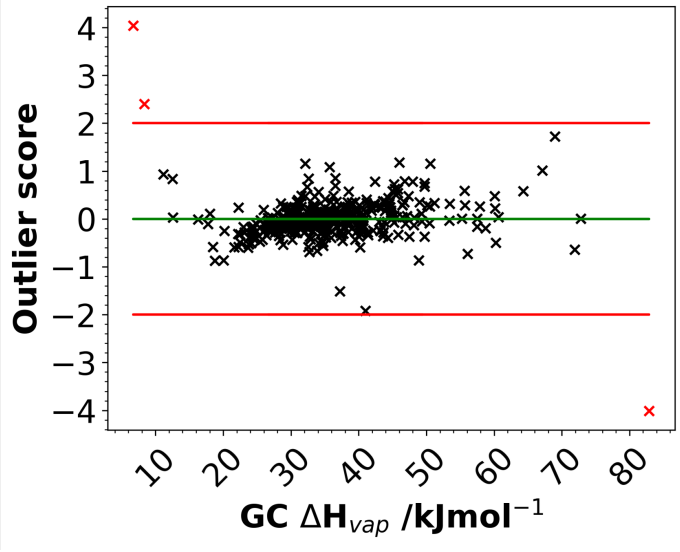


Fig. S5 ΔH_{vap} outlier analysis and scores after removing erroneous data from the CRC Handbook of Chemistry and Physics

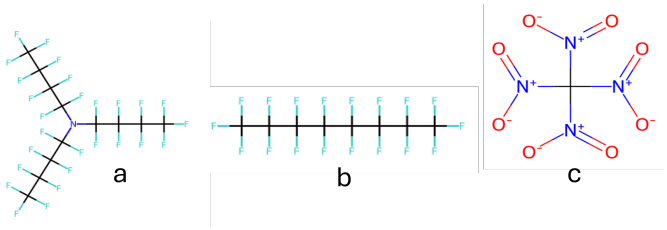


Fig. S6 Molecules with large bias in JR GC ΔH_{vap} predictions

fixed at values close to what may be expected as the average 95 % confidence interval uncertainty associated with experimental measurements of T_m —options b and c —, we observe signs of overfitting as the model fits the training set data almost perfectly but provides inferior generalization for unseen data compared to options b and c.

S2.3 Out of Sample Mean Prediction

Table S4 provides numerical results comparing GC-only and GCGP predictions of ΔH_{vap} for highly fluorinated molecules.

Table S4 Comparison of GCGP and JR GC ΔH_{vap} predictions for highly fluorinated molecules. GCGP Unc. represent GP predicted 95 % confidence intervals magnitude on the property predictions. Experimental data from Yaws' Critical Property Data for Chemical Engineers and Chemists[3]. * = $kJmol^{-1}$

Molecular Formula	CAS No.	MW $/gmol^{-1}$	JR GC ΔH_{vap} /*	Exp. ΔH_{vap} /*	GCGP ΔH_{vap} /*	GCGP Unc. /*
$C_{10}F_{18}$	306-94-5	462.08	9.68	35.80	35.6	9.7
C_9F_{20}	375-96-2	488.07	7.62	34.22	35.3	14.0
C_6F_{14}	355-04-4	338.05	9.74	27.89	25.8	16.7
$C_8HF_{15}O_2$	335-67-1	414.07	29.24	40.40	42.1	20.0
$C_9HF_{17}O_2$	375-95-1	464.08	28.54	43.03	45.2	22.1

We see a remarkable correction of the systematic bias in the GC-only prediction of ΔH_{vap} for highly fluorinated molecules. High uncertainty is as a result of the very limited data for highly fluorinated molecules present in the training set.

S2.4 GCGP is Robust to Model/Kernel Choices

S2.4.1 Uncertainty in LML

The feature-based stratified splitting algorithm is designed to be robust to random seed choice. Only ΔH_{vap} and T_m gave different train/test splits for different random seed values. For ΔH_{vap} and T_m, we generated an ensemble of ten training sets from the data sets using 10 (randomly generated) random seeds in addition to random seed 42 used in the final implementation. We computed LML values and key model performance metrics for the final model/kernel combination in this work. We used the computed LML values to obtain a rough estimate of what may be considered a reasonable lower bound on the uncertainty in LML values for our work. Table S6 summarizes these results. Based on data in Table S6, we consider a 1 % relative error in the LML values as a reasonable lower bound on the average uncertainties associated with the final reported LML values. LML values within a 1 % difference from each other are considered similar.

S2.4.2 Comparisons of Kernels

Table S5 shows a comparison of isotropic kernels for Model 3 for all properties.

Table S5 Kernel rankings for isotropic kernels with Model 3. Lowest total is best

	Mt12	Mt32	Mt52	RBF	**RQ**
T_m	1	1	1	5	1
T_b	1	2	2	5	2
T_c	1	2	4	5	3
V_c	1	3	3	5	1
P_c	5	4	3	1	1
H_{vap}	1	2	3	5	2
Total	10	14	16	26	10

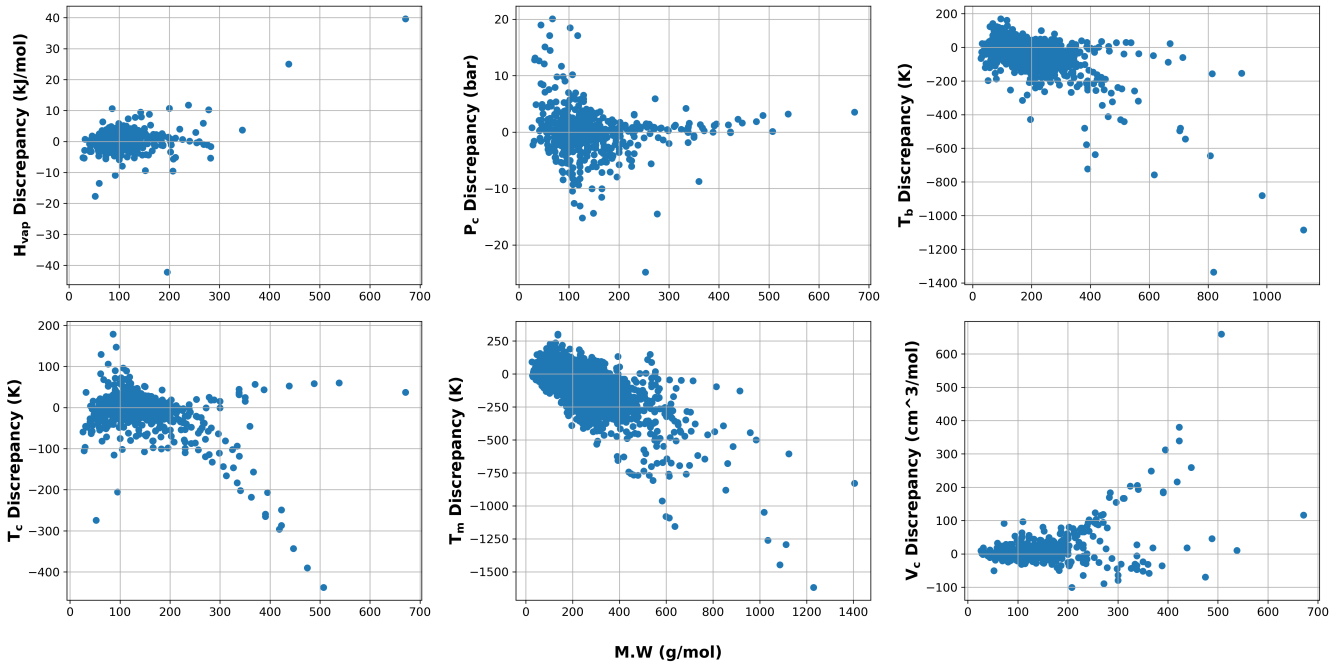


Fig. S7 Data visualization for all properties of interest. For each subplot, the x-axis is the molecular weight, the y-axis is the discrepancy between the experimental and GC predicted property value.

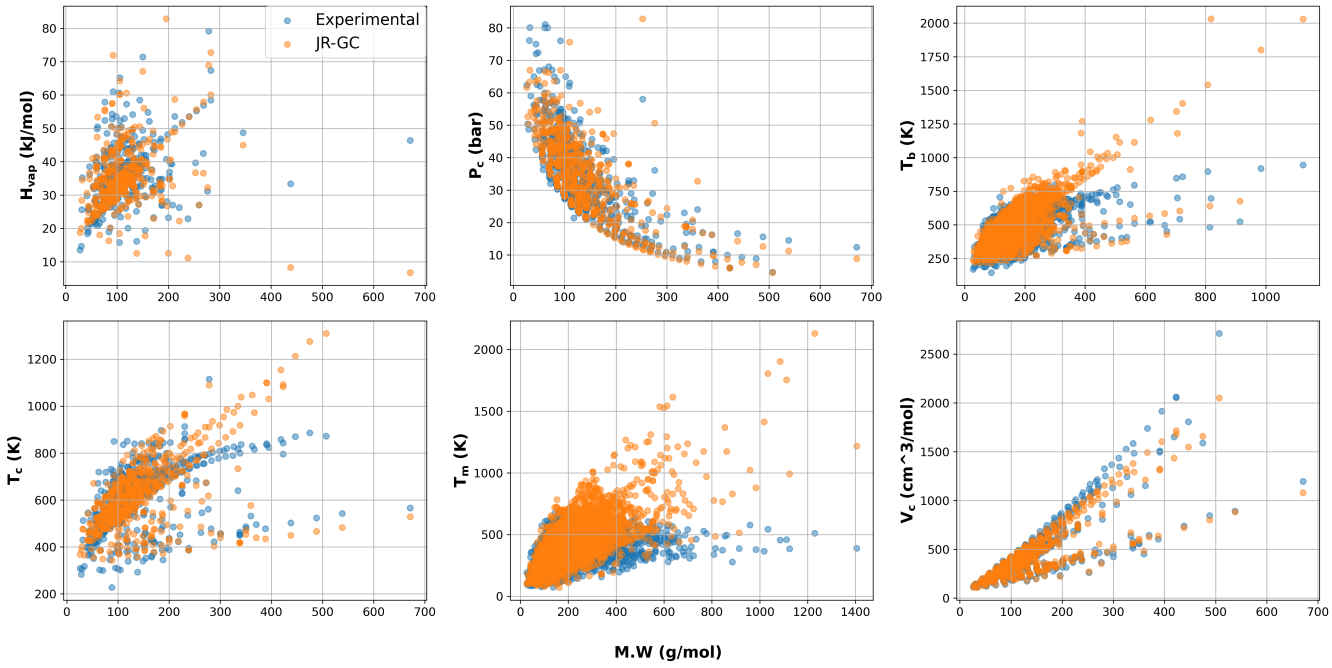


Fig. S8 Data visualization for all properties of interest. For each subplot, the x-axis is the molecular weight, the y-axis is the experimental property value (blue) or GC predicted property value (orange).

The numbers represent the rank of the kernel based on LML values considering 1 % differences as the criterion to judge that one kernel has a better LML than another for a given property. The RQ kernel and Matern kernel with $\nu = 1/2$ have the lowest overall sum of ranks (best score) with a value of 10, followed by the Matern kernels with ν of 3/2 and 5/2. The RQ kernel has a shape parameter also known as scale-mixture parameter α, which makes it flexible for use in modeling a variety of data with different or unclear smoothness trends. This also makes the RQ kernel a more physically and mathematically justifiable kernel choice for modeling all properties in this work. The Matérn kernel with $\nu = 1/2$, showed excellent performance for most properties but was the worst performing for P_c. This is likely due to the Matérn with $\nu = 1/2$ kernel's intrinsic property of modeling func-

tions that are continuous but not differentiable, which is poorly suited for capturing the higher degree of smoothness characteristic of the underlying functional mapping for P_c. The RQ kernel is a general-purpose choice that provides competitive performance across all properties modeled in this work.

S2.5 Effects of Train/Test Splits

Table S6 shows that the key perfromance metrics of the GCGP method with stratified splitting are robust to the choice of random seed used for determining train/test splits. Furthermore, the train set metrics are generally slightly better than those of the testing set as expected for well trained models with good transferability to unseen data.

Table S6 Effects of Random Seed Choice for Train/Test Splits on Model Performance Metrics for Training and Testing Sets. Seed 42 was used for generating all final results in the main text. * means kJ/mol or K for ΔH_{vap} or T_m respectively.

Property	Seed	LML	Test MAPE %	Train MAPE %	Test MAE *	Train MAE *	Test RMSE *	Train RMSE *	Test R^2	Train R^2
ΔH_{vap}	42	-148.38	4.45	4.58	1.62	1.61	2.42	2.38	0.93	0.92
	670487	-102.01	5.79	4.18	1.97	1.49	3.33	2.13	0.85	0.93
	116739	-115.87	4.88	4.41	1.59	1.58	2.61	2.25	0.88	0.93
	26225	-135.43	4.18	4.53	1.44	1.61	2.03	2.35	0.93	0.92
	777572	-110.25	5.56	4.14	1.99	1.46	3.18	2.11	0.87	0.93
	288389	-120.42	5.42	4.37	1.87	1.54	3.2	2.23	0.89	0.92
	256787	-108.68	5.05	4.38	1.68	1.57	2.91	2.21	0.86	0.93
	234053	-110.38	5.39	4.2	1.78	1.5	2.8	2.15	0.88	0.93
	146316	-96.82	6.41	4.04	2.08	1.46	3.51	2.08	0.85	0.94
	772246	-139.44	4.62	4.59	1.72	1.61	2.7	2.37	0.92	0.92
	107473	-115.61	5.52	4.27	1.76	1.54	2.77	2.25	0.87	0.93
T_m	42	-3417.48	12.36	12.55	40.57	39.66	53.41	52.71	0.73	0.75
	670487	-3427.2	12.5	12.56	40.86	39.67	54.3	52.57	0.73	0.75
	116739	-3428.64	12.05	12.6	39.42	39.81	53.12	52.68	0.74	0.75
	26225	-3431.1	12.39	12.55	39.84	39.8	53.11	52.73	0.74	0.75
	777572	-3460.69	12.27	12.62	39.45	40.01	51.83	53.09	0.75	0.74
	288389	-3401.59	12.66	12.45	40.49	39.53	54.01	52.39	0.72	0.75
	256787	-3423.49	12.14	12.62	40.03	39.78	53.24	52.78	0.73	0.75
	234053	-3420.36	12.58	12.44	40.7	39.41	53.61	52.4	0.74	0.75
	146316	-3447.06	12.14	12.55	39.96	39.69	53.79	52.57	0.74	0.74
	772246	-3441.01	12.46	12.48	40.66	39.5	53.89	52.52	0.74	0.75
	107473	-3453.46	12.18	12.58	40.06	39.69	52.87	52.79	0.75	0.74

Notes and references

1 K. Joback and R. Reid, *Chem. Eng. Commun.*, 1987, **57**, 233–243.

2 J. R. Rumble, *CRC Handbook of Chemistry and Physics*, CRC Press/Taylor & Francis, Boca Raton, FL, 105th edn, 2023.

3 C. L. Yaws, *Yaws' Critical Property Data for Chemical Engineers and Chemists*, Knovel, 2014.

4 National Institute of Standards and Technology, Office of Data, *NIST Chemistry WebBook*, https://webbook.nist.gov/chemistry/, 2024.

5 K. B. Wiberg and R. F. Waldron, *J. Am. Chem. Soc.*, 1991, **113**, 7697–7705.

6 I. I. Vasil'eva, A. A. Naumova, A. A. Polyakov, T. N. Tyvina and N. V. Kozlova, *Zh. Prikl. Khim. (Leningrad)*, 1990, **63**, 1879–1881.

7 R. M. Stephenson and S. Malanowski, *Handbook of the Thermodynamics of Organic Compounds*, Springer Netherlands, Dordrecht, 1987.

8 *CRC Handbook of Data on Organic Compounds*, ed. R. C. Weast and J. G. Grasselli, CRC Press, Inc., Boca Raton, FL, 2nd edn, 1989, vol. 1.

9 A. van Roon, J. R. Parsons and H. A. J. Govers, *J. Chromatogr. A*, 2002, **955**, 105–115.