Supporting Information for:

Active learning for efficient navigation of multi-component gas adsorption landscapes in a MOF

Krishnendu Mukherjee,[1] Etinosa Osaro,[1] and Yamil J. Colón*[1]

[1]Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, IN, 46556, USA
*Corresponding author: ycolon@nd.edu

GP Kernel selection process

Here is the performance metric to compare different kernel combination's performance and find the best kernel/combination of kernels for the GPs. This metric is calculated after 500 iterations of Active learning using the following equation.

$$\text{Performance} = \frac{\overline{100 - AC_1} + \overline{100 - AC_2} + \overline{AC_{CL}} + \overline{MRE_1} + \overline{MRE_2}}{5} \quad (1)$$

This parameter (shown in the y-axis of figure S1, S3, and S7-S9) is a lumped-variable for five different parameters. Each of these five parameters for the kernel shows a desired performance from the Active learning protocol. A low value of all these parameters are desirable. Also, each term in the above equation in performance calculation corresponds to the mean of accuracy over 500 iterations, not just a point at 500 iterations. This is done to find the cumulative performance of the kernel combinations.

The first two terms shows absolute distance of the accuracy from the desired value of 100% (for the dual GPs) for gas species 1 and 2 in the mixture. Further the distance parameter is normalized with maximum and minimum values, so that final parameter is within 0 and 1.

$$\overline{100 - AC_i} = \left| \frac{(100 - AC_i) - min(100 - AC_i)}{max(100 - AC_i) - min(100 - AC_i)} \right| \quad (2)$$

The third parameter is the difference between accuracies of the species of C and L normalized by maximum and minimum value of the parameter.

$$\overline{AC_{CL}} = \left| \frac{(|AC_C - AC_L|) - min(|AC_C - AC_L|)}{max(|AC_C - AC_L|) - min(|AC_C - AC_L|)} \right| \quad (3)$$

The fourth and the fifth shows the mean relative errors. Also, the first three accuracy-based parameters are normalized with respect to maximum and minimum values, while the MRE-based parameters are scaled with maximum value only. This is done to because many kernels were too close to the lowest MREs and hence they would be very close to 0 if we had used the min-max scaling. We just scaled it with respect to maximum value so that the MRE distribution remains comparable with accuracy distribution in the visualization.

$$\overline{MRE_i} = \frac{MRE_i}{max(MRE_i)} \quad (4)$$

Thus, we get a final performance-based parameter, which tells us a how well a kernel combination is performing with respect to other combinations. Therefore, the final parameter encapsulates whether the GPs are close to 100% accuracy, how GPs accuracies are close to one another, and how low the MREs are (all of them are given an equal weight). We select the kernel/kernel combination with the lowest value of this performance metric.

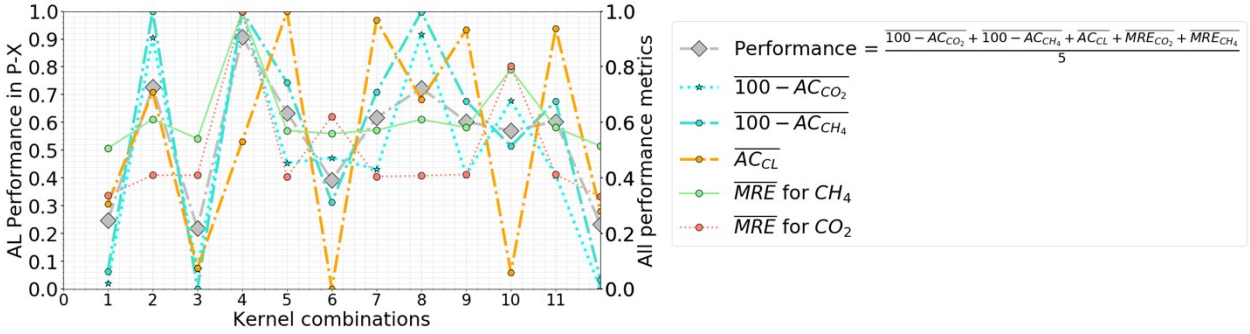

Figure S1. Performance plots for kernel optimization for CO_2-CH_4 mixture after 500 iterations of active learning (all the parameters are averaged over all the iterations leading to 500th iteration). The x-axis shows the kernel index combination. For reference, indices 1, 2 and 3 are: rational quadratic (RQ), Matérn (M), and radial basis function (RBF). The next indices (from 4 to 12) follows the combination of either of these two but not three at a time. Index 4 would be RQ + RQ, index 5 RQ + M, 6 would be RQ + RBF, and so on. The y-axis on the right shows the performance metrics, which encapsulates five entities from the dual-GPs. First two terms quantify how close the accuracies of kernels to 100. The third, how close the two accuracies are to one another. Fourth and fifth ones finds the mean relative errors (MREs) for both the GPs. In essence, the optimization is done to find the kernel combination which gives the minimum of this lumped-performance parameter. Here, we find the Index 3 (RBF) to be the best in the performance, followed closely by dual-RBF at Index 12, and then Index 1 by a single RQ.

Figure S1 shows the performance of all kernels for CO_2-CH_4 mixture. We find that Index 3 (a single RBF) outperforms all the other combinations. Hence, we go forward with RBF for the final fit. Also, for the kernel optimizations of Xe-Kr and H_2S-CO_2 systems, we found RQ (Index 1) to be the best performer. It is interesting to note that single kernels performed well and particularly RBF which has only two parameters. In figure S1, the RBF and RQ kernels have the lowest value of the parameters, $|100 - \overline{AC_{CO_2}}|$ and $|100 - \overline{AC_{CH_4}}|$. This shows these kernels candidates showed a higher accuracy after 500 iterations. Their MREs were just marginally lower than other combination, however their accuracies were much closer which provided a boost to the performance. We have added the kernel optimization results for the other two mixtures in the figure S3.

RASPA Input file for GCMC simulations

- The values of features P, X_1/X_2, and T were changed according with the states
- The molecules names and definition corresponds to the RASPA forcefield files and folder location. The same input file was used for all the three mixtures, only the adsorbate forcefield was changed in simulation.

SimulationType MonteCarlo

NumberOfCycles 50000

NumberOfInitializationCycles 5000

PrintEvery 1000

ContinueAfterCrash no

WriteBinaryRestartFileEvery 2000

UseChargesFromCIFFile yes

Forcefield GenericMOFs

RemoveAtomNumberCodeFromLabel yes

Framework 0

FrameworkName Cu-BTC

UnitCells 1 1 1

ExternalTemperature T

ExternalPressure P

Component 0 MoleculeName CO2/H2S/Xe

MoleculeDefinition TraPPE

MolFraction X_1

TranslationProbability 0.5

RotationProbability 0.5

ReinsertionProbability 0.5

RegrowProbability 0.5

IdentityChangeProbability 1.0

NumberOfIdentityChanges 2

IdentityChangesList 0 1

SwapProbability 1.0

CreateNumberOfMolecules 0

Component 1 MoleculeName methane/CO2/Kr

MoleculeDefinition TraPPE

MolFraction X_2

TranslationProbability 0.5

ReinsertionProbability 0.5

RegrowProbability 0.5

RotationProbability 0.5

IdentityChangeProbability 1.0

NumberOfIdentityChanges 2

IdentityChangesList 0 1

SwapProbability 1.0

CreateNumberOfMolecules 0

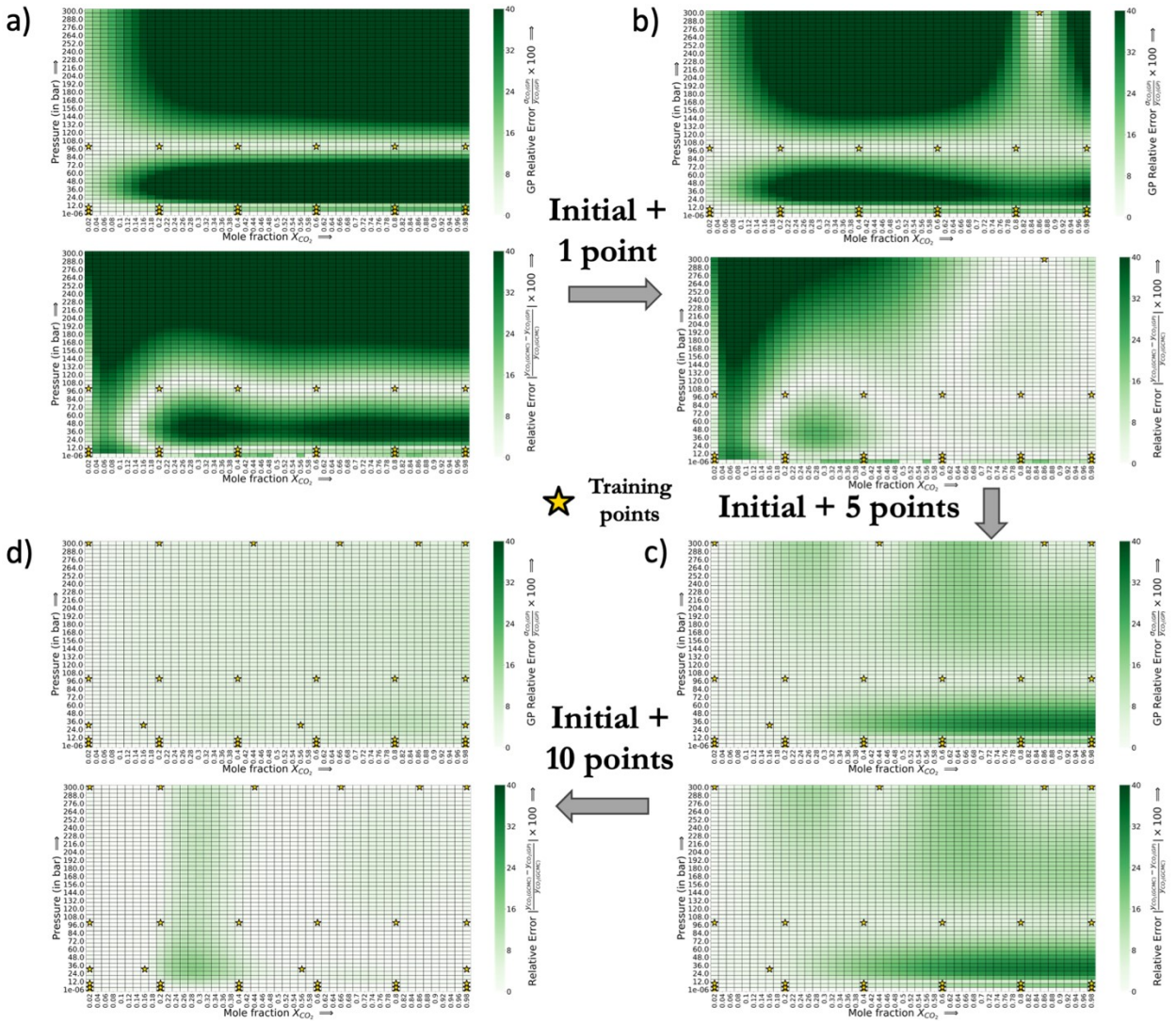

Figure S2. Error maps for GP relative error and the absolute relative error (true error on comparison with GCMC) for CO_2 in the CO_2-CH_4 mixture at different stages of AL. a) Trained only on the initial dataset, b) Initial dataset + 1 point, c) Initial + 5 points, d) Initial + 10 points. We observe that as more points are provided to the GPs, both the GP's perceived error and the relative error starts to look similar. The AL only up to 10 points is shown. For the case of CO_2-CH_4, the AL goes to add 21 extra point besides the initial training set. Here only till 10 points of AL is shown to highlight the changes in the error as new points based on their uncertainty are provided to the GP.

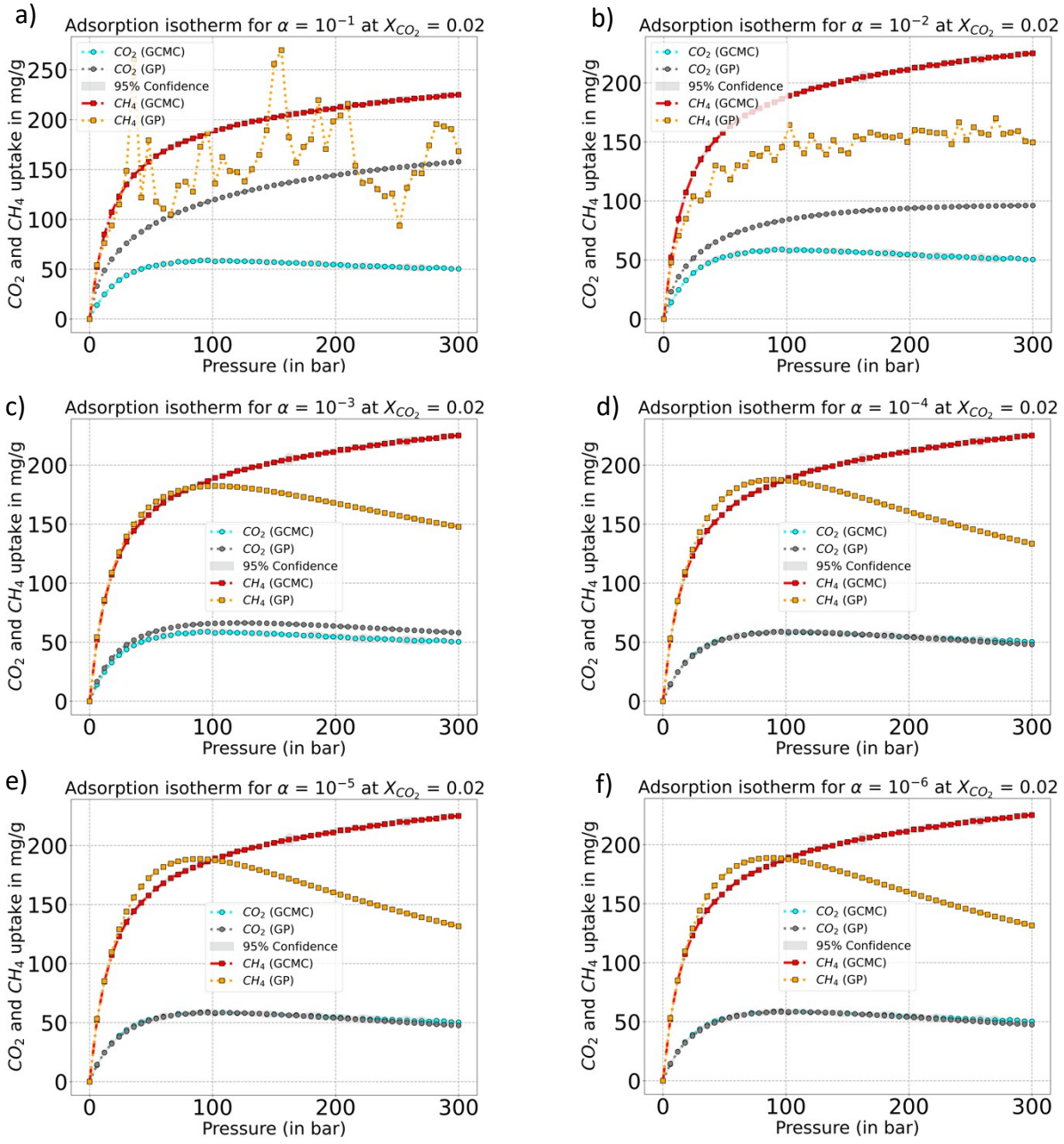

Figure S3. Comparing the adsorption isotherm plots for CO_2–CH_4 mixtures at X_{CO2} = 0.02 at 300K for different regularization parameter α (for a single RQ kernel). The plot consists of α of a) 10^{-1}, b) 10^{-2}, c) 10^{-3}, d) 10^{-4}, e) 10^{-5}, and f) 10^{-6}. The goal was to avoid high-fluctuation which can results from a high α, which can provide a high-threshold for variance. We observe that fluctuations cease to exist at α of 10^{-3}. The next consideration was a balanced fit between the GP and the GCMC adsorption data. We find at α of 10^{-4} gives a better fit for both CO_2 and CH_4. Hence, we keep the value for this gas mixture for both the P–X and P–X–T phase space active learning.

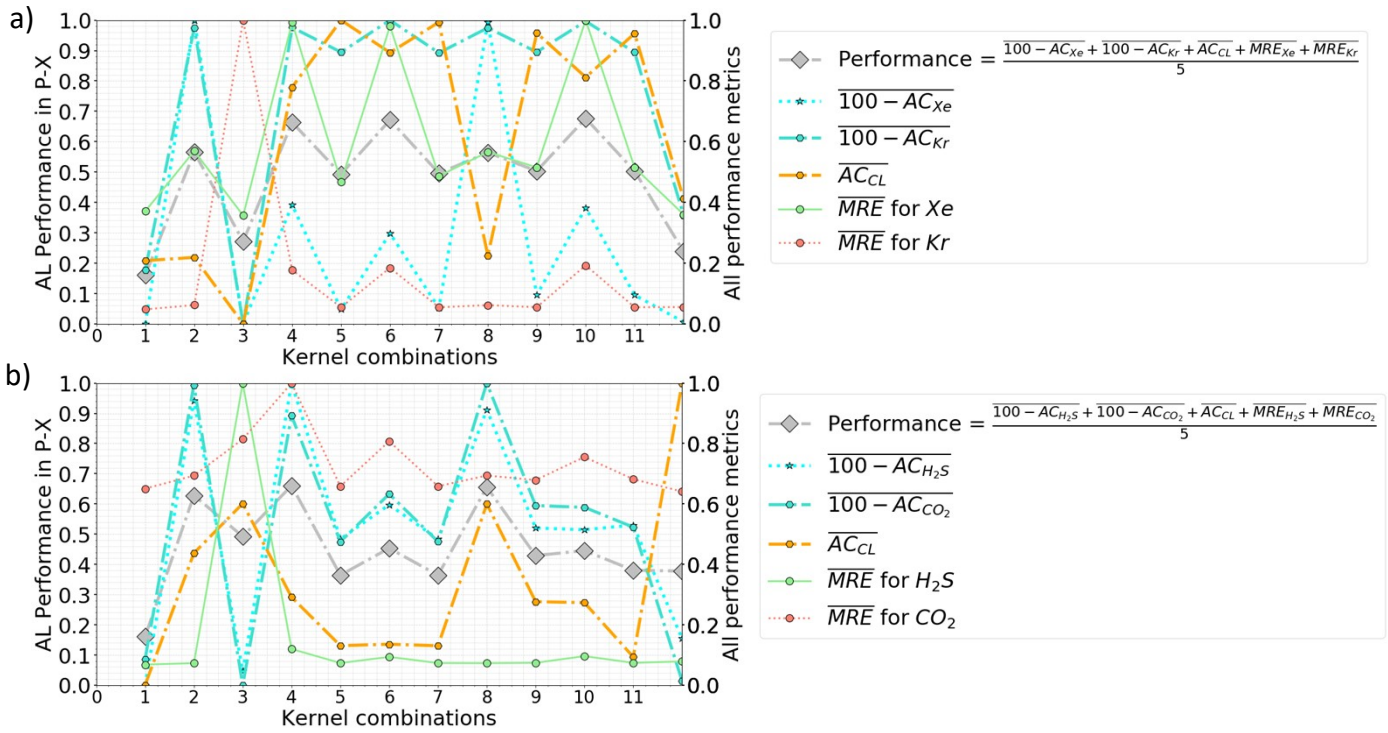

Figure S4. Kernel optimization results after 500 iterations of Active learning in the P–X phase space for a) Xe–Kr and b) H_2S–CO_2. The performance parameter in left side y-axis is a lumped parameter for five variables which are listed in the legends. From both a) and b), we observe that the index 1 is performing the best for both these systems. In figure 2, we also observe similar outcome for CO_2–CH_4. Thus, Rational Quadratic (RQ) performs best for all the three systems.

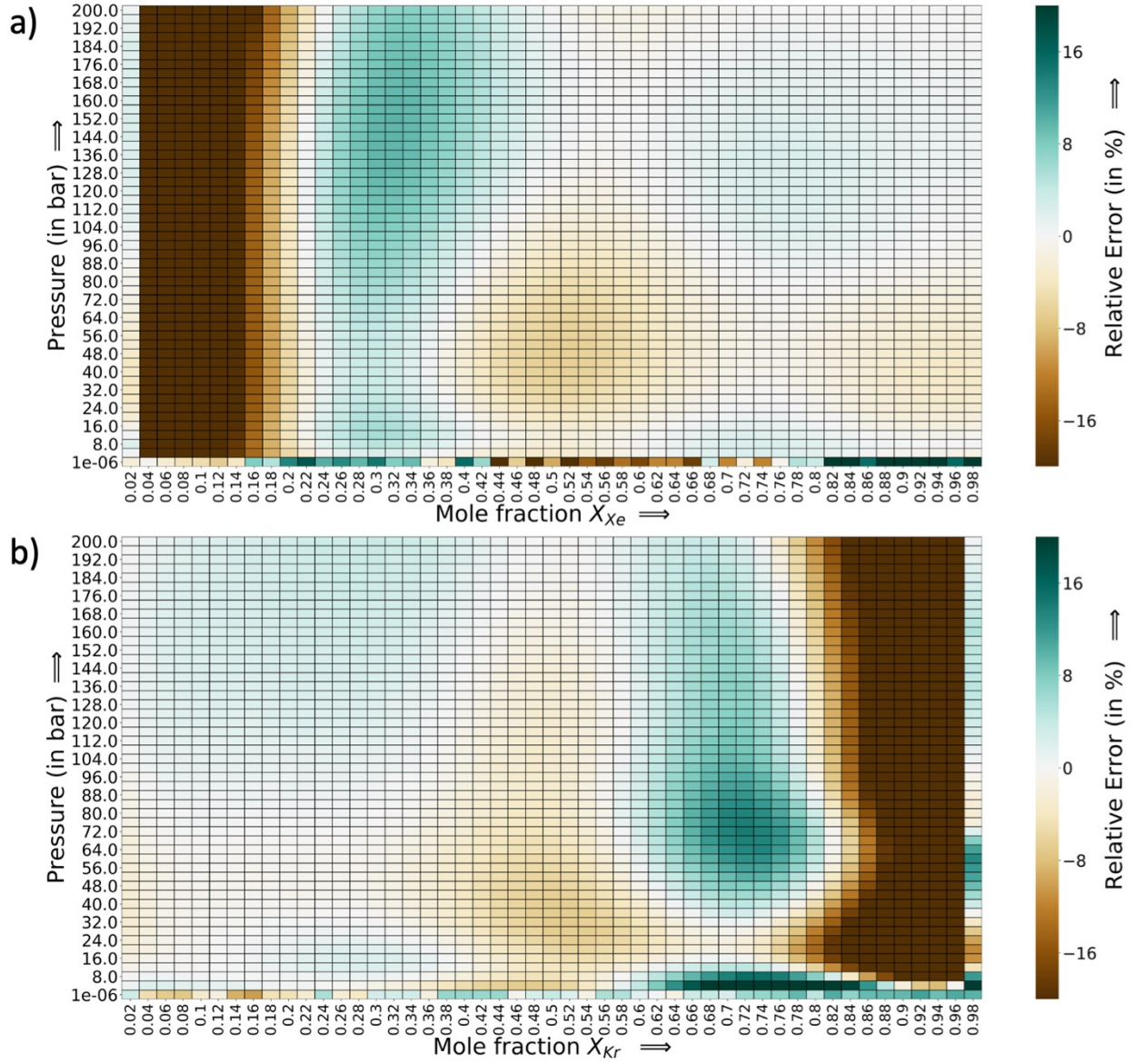

Figure S5. Error heat map for Xe–Kr system at the 90% PAC cut-off for a) Xe, and b) Kr. We find the densest region of high errors lies at low Xe-high Kr range. This again corresponded to the component that is more attractive to the Cu-BTC. Also, the errors are more under-prediction, only a small section corresponds to over-prediction.

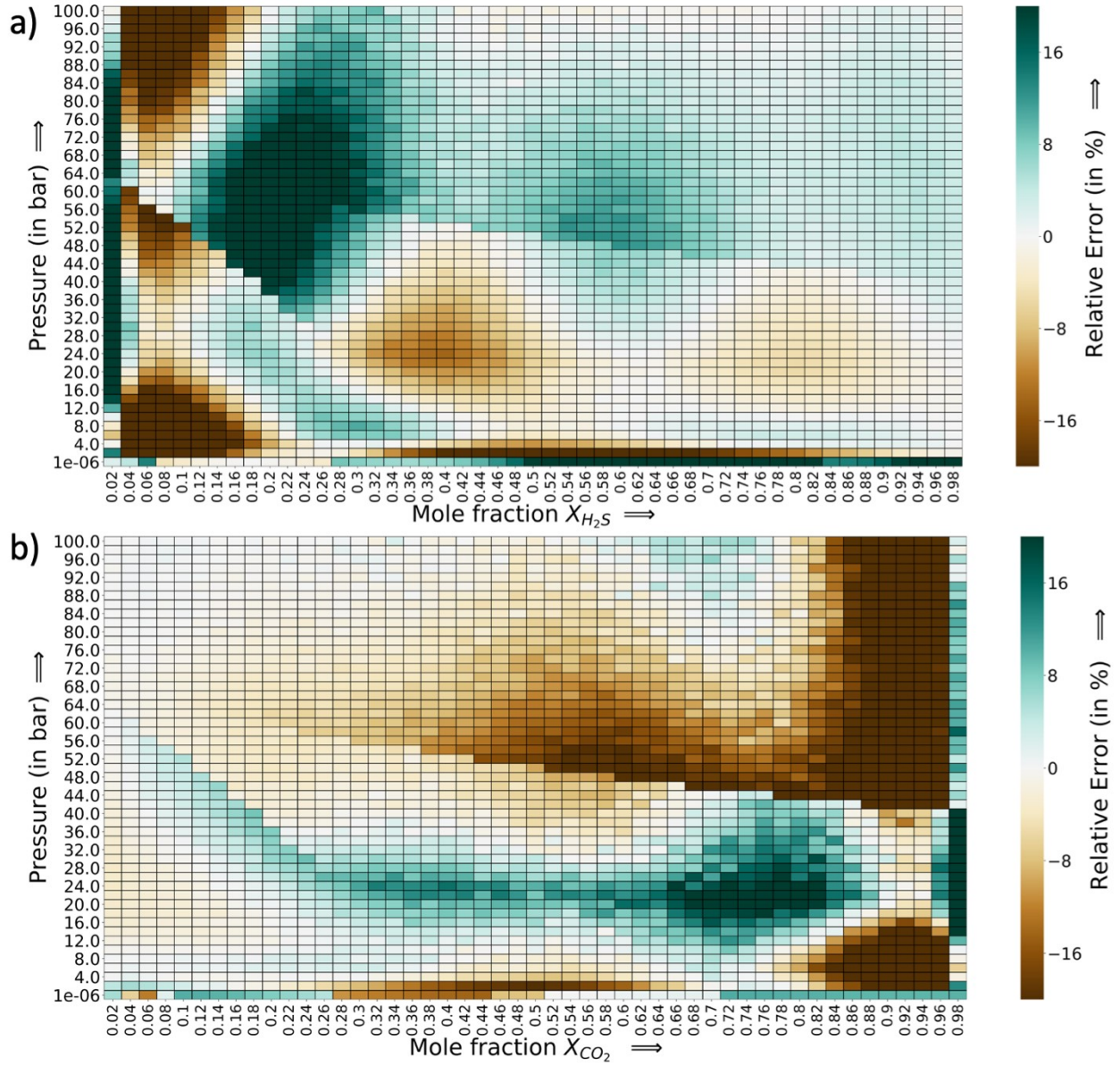

Figure S6. Error heat map for H_2S–CO_2 system at the 90% PAC cut-off for a) H_2S, and b) CO_2. We find the densest region of high errors lies at low H_2S-high CO_2 range.

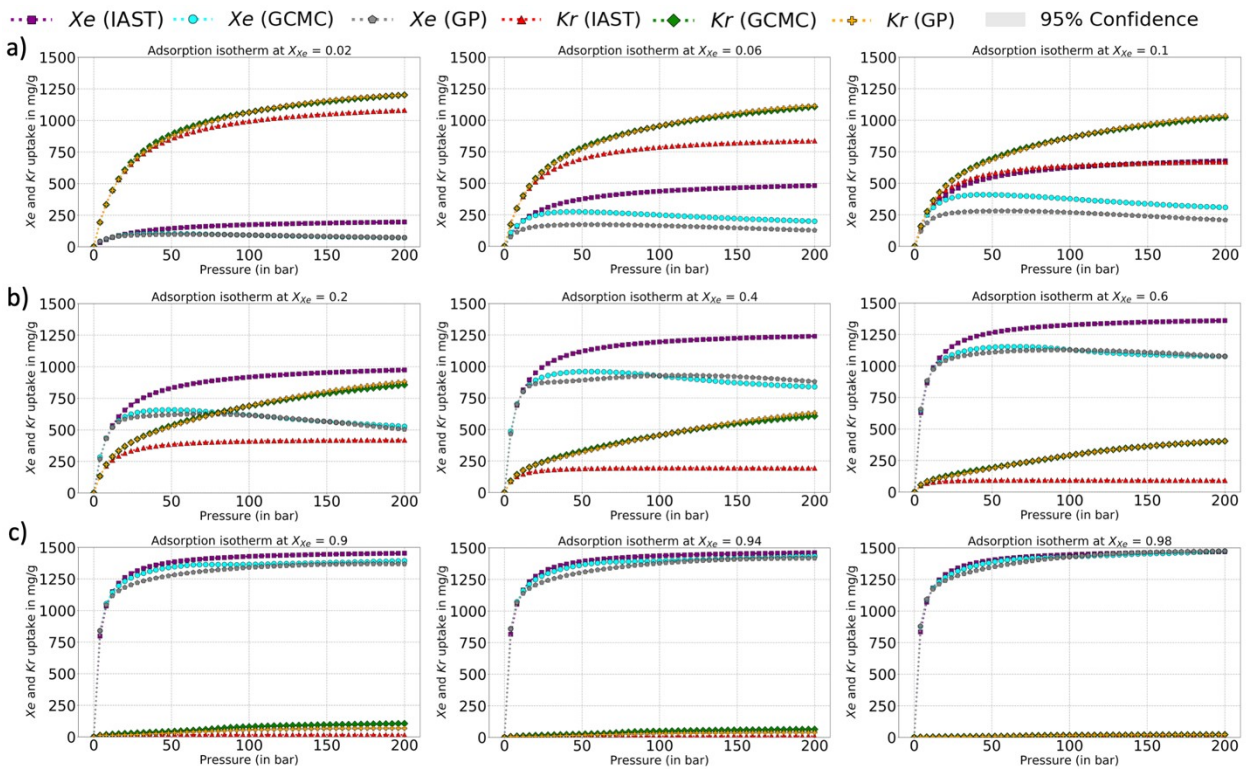

Figure S7. Adsorption plots for Xe–Kr system (RQ kernel, Index 1 of figure S2a) for the P-X phase space. a) X_{Xe} = 0.02, 0.06 and 0.10, b) X_{Xe} = 0.20, 0.40 and 0.60, and c) X_{Xe} = 0.90, 0.94 and 0.98. The highest deviation of the GP-predicted adsorption is seen for Xe at the low concentration of X_{Xe}. The GP-predicted adsorption matches well with GCMC beyond this range. Also, GP-model accurately captures Kr uptake for nearly all the pressure and temperature range.

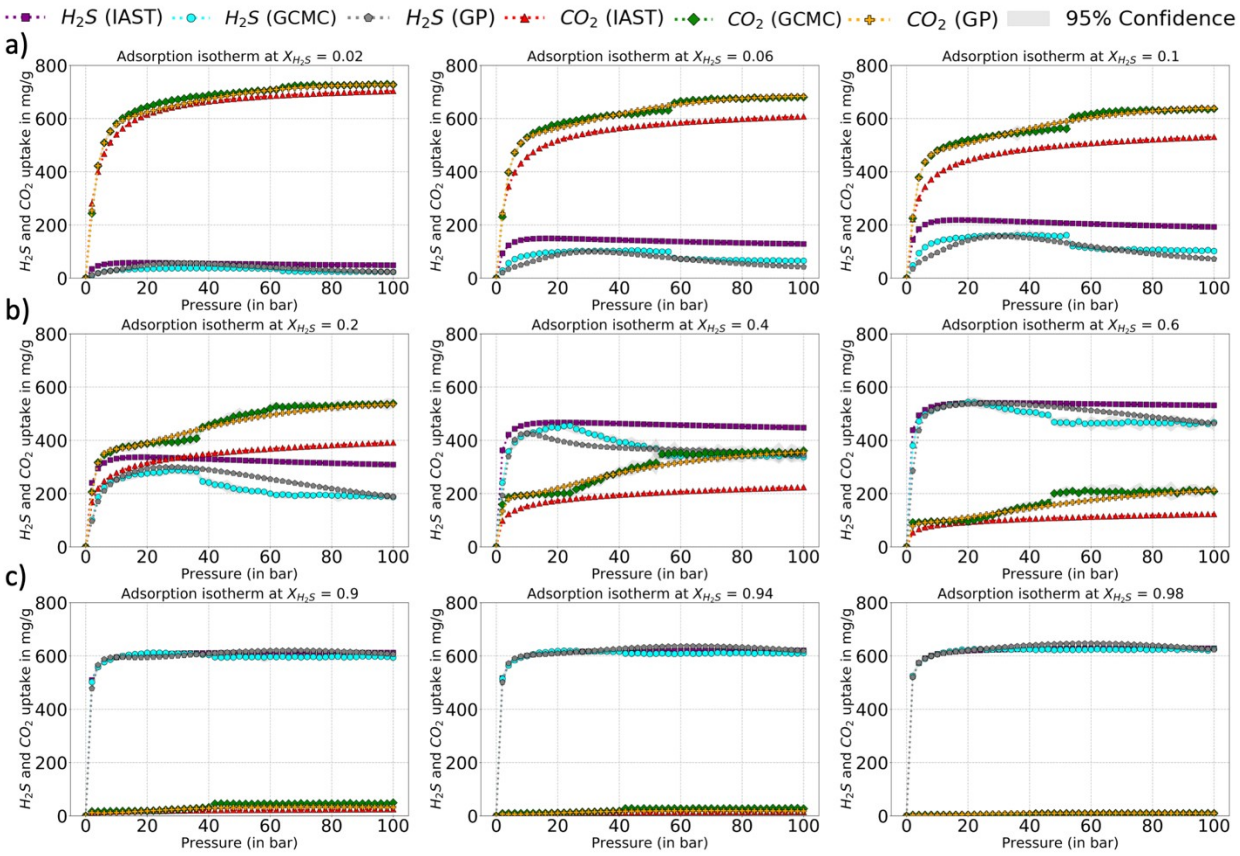

Figure S8. Adsorption plots for H_2S–CO_2 system (RQ kernel, Index 1 of figure S2b) for the P-X phase space. a) X_{H_2S} = 0.02, 0.06 and 0.10, b) X_{H_2S} = 0.20, 0.40 and 0.60, and c) X_{H_2S} = 0.90, 0.94 and 0.98. The highest deviation of the GP-predicted adsorption is seen for H_2S at medium concentration of X_{H_2S}. The GP-predicted adsorption matches well with GCMC at low and high values of X_{H_2S}. Also, GP-model accurately captures CO_2 uptake on the low and high concentration of X_{H_2S}. However, the errors are higher for CO2 in medium concentration of X_{H_2S}, though less compared to that of H_2S.

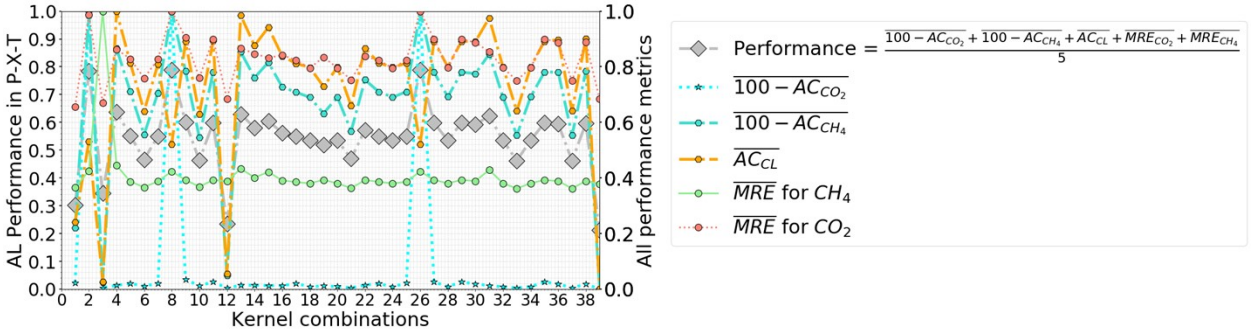

Figure S9. Performance plots for kernel optimization for CO_2-CH_4 mixture after 500 iterations of active learning for P-X-T feature space. Here again there are 39 candidate kernel combinations. Like P-X feature, indices 1, 2 and 3 are: rational quadratic (RQ), Matérn (M), and radial basis function (RBF), and the rest of the combinations follows adding the kernels in that sequence only. Here, we find the best kernel to be index 39, which is RBF + RBF + RBF, followed very closely by kernel 12 which is dual RBF, followed by index 1 (RQ), and then index 3 (RBF). With these observations, RBF kernels for the GP provides a good fit for the CO_2-CH_4 mixture.

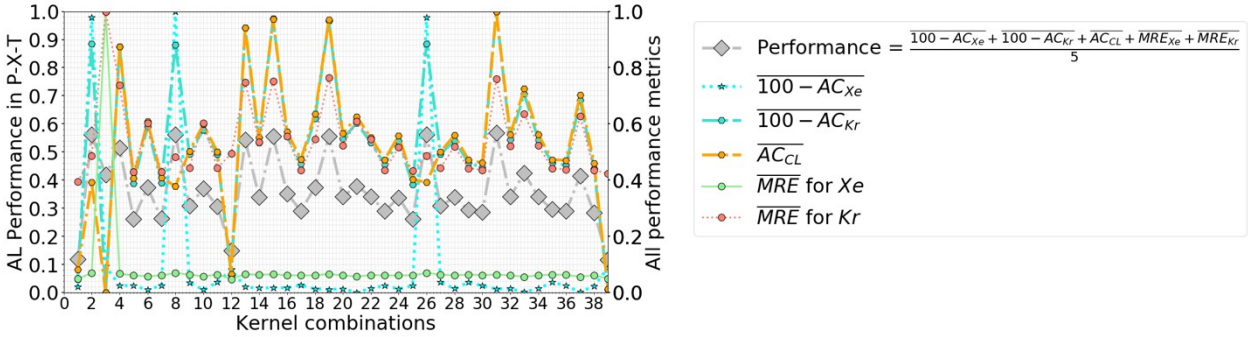

Figure S10. Performance plots for kernel optimization for Xe-Kr mixture after 500 iterations of active learning for P-X-T phase space. The best fit here corresponds to index 39 (triple RBF with performance: 0.1158), followed very closely by index 1 (RQ with performance: 0.1179).

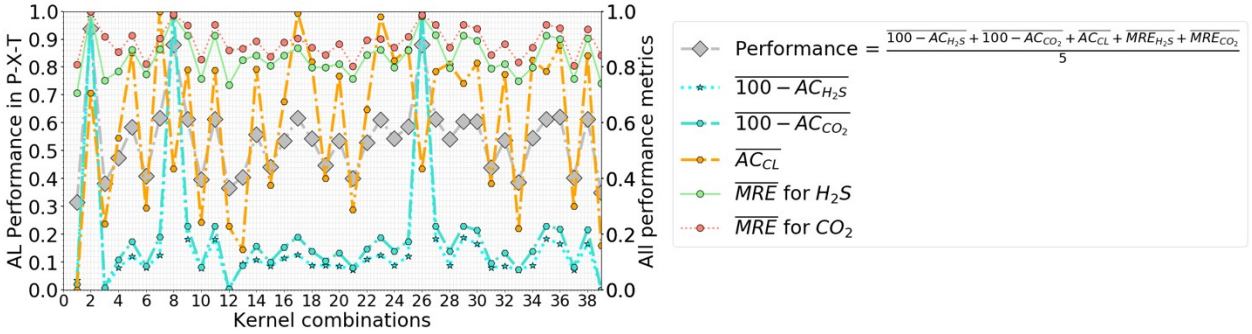

Figure S11. Performance plots for kernel optimization for H_2S-CO_2 mixture after 500 iterations of active learning for P-X-T phase space. The best fit here corresponds to index 1 (RQ), followed by index 39 (triple RBF).

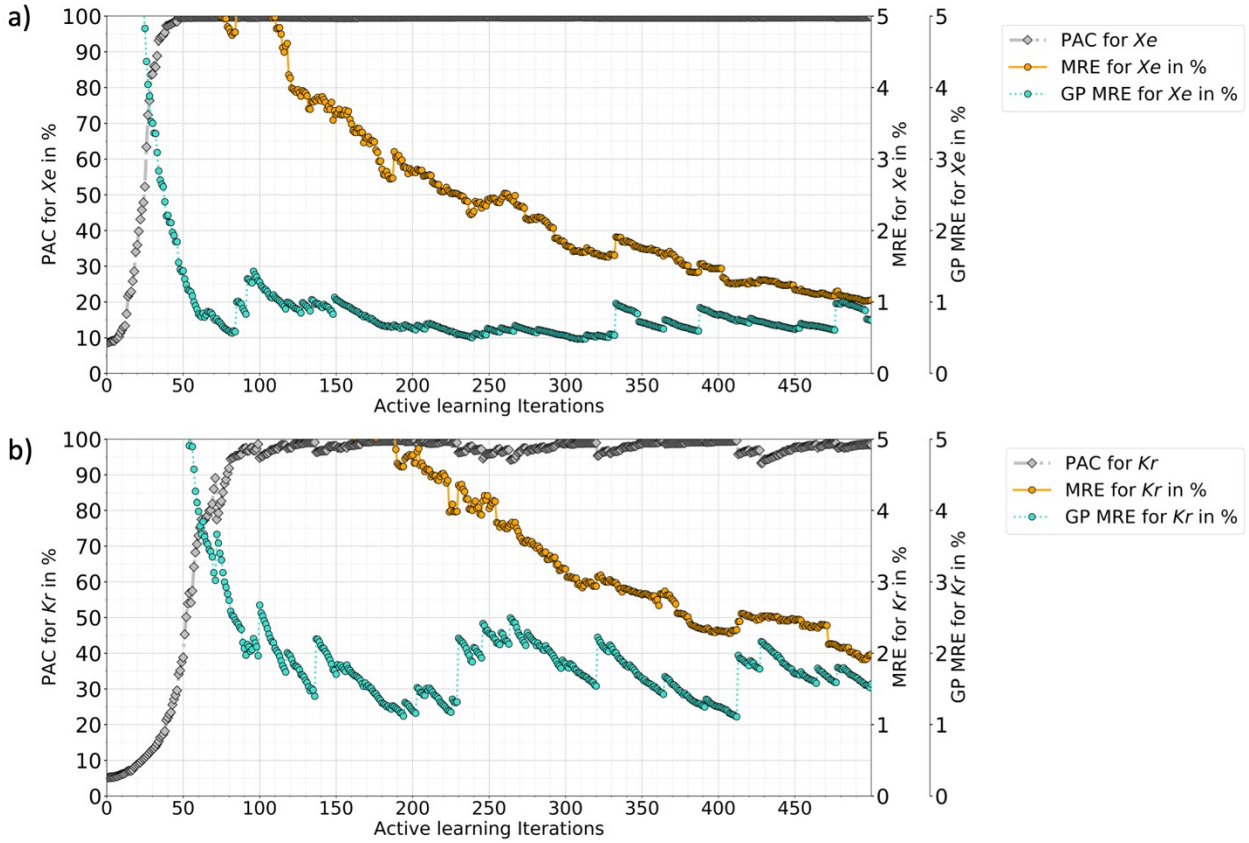

Figure S12. Active learning progression plots for a single RBF + RBF + RBF kernel for P-X-T phase space a) Xe, and b) Kr. As we see with subsequent iterations of the learning, the MRE (true error with respect to the ground truth) converges with the GP MRE. However, as we had set the accuracy threshold of 90%, the AL process will finish much earlier for a desired performance. This is just to show if the AL was to progress how the performance would be in the following iterations.

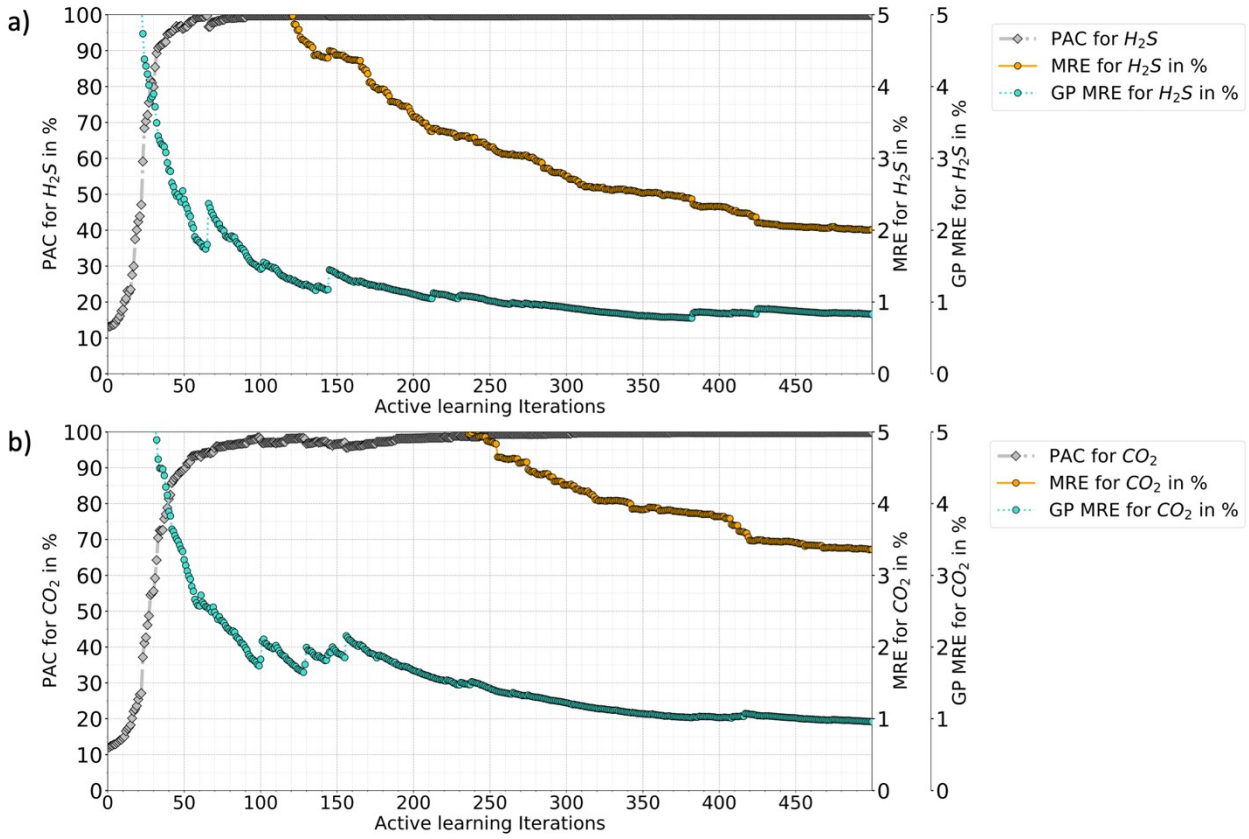

Figure S13. Active learning progression plots for a single RQ kernel for P-X-T phase space a) H_2S, and b) CO_2.

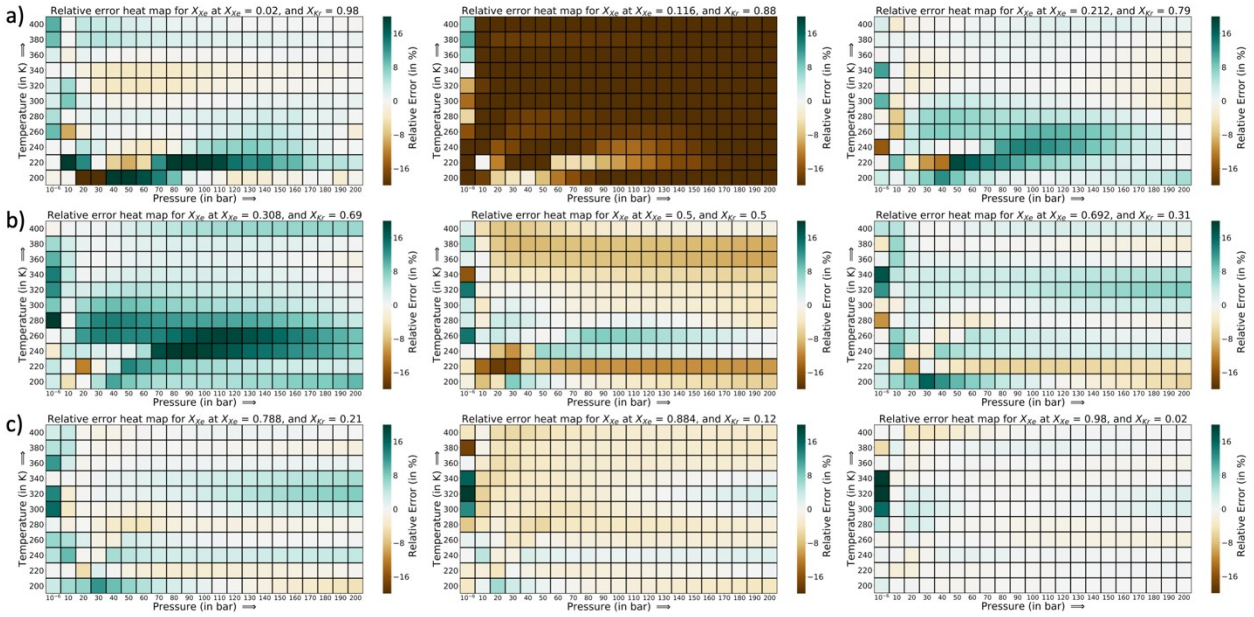

Figure S14. Relative error heat maps at the 90% PAC cut-off for Xe in the Xe-Kr mixture with triple-RBF kernel, a) X_{Xe} = 0.02, 0. 116, and 0.212, b) X_{Xe} = 0.308, 0.5, and 0.692, and c) X_{Xe} = 0.788, 0.884, and 0.98. We find the region of X_{Xe} = 0.116 (X_{Kr} = 0.884) having the highest errors for Xe uptake, with most errors showing that GP model in under-predicting. After this region, the error regions become less dense, with slight over-prediction by the GP for Xe compared to GCMC.

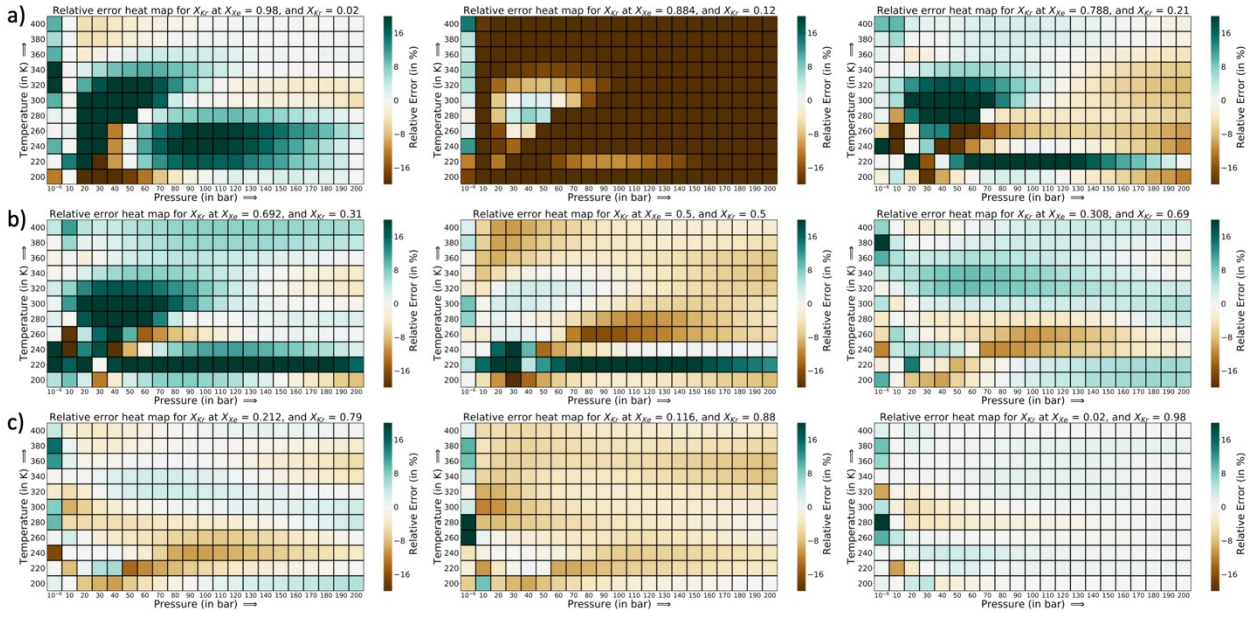

Figure S15. Relative error heat maps at the 90% PAC cut-off for Kr in the Xe-Kr mixture with triple-RBF kernel, a) X_{Kr} = 0.02, 0. 116, and 0.212, b) X_{Kr} = 0.308, 0.5, and 0.692, and c) X_{Kr} = 0.788, 0.884, and 0.98. We find the region of X_{Kr} = 0.116 (X_{Xe} = 0.884) having the highest errors for Kr uptake.

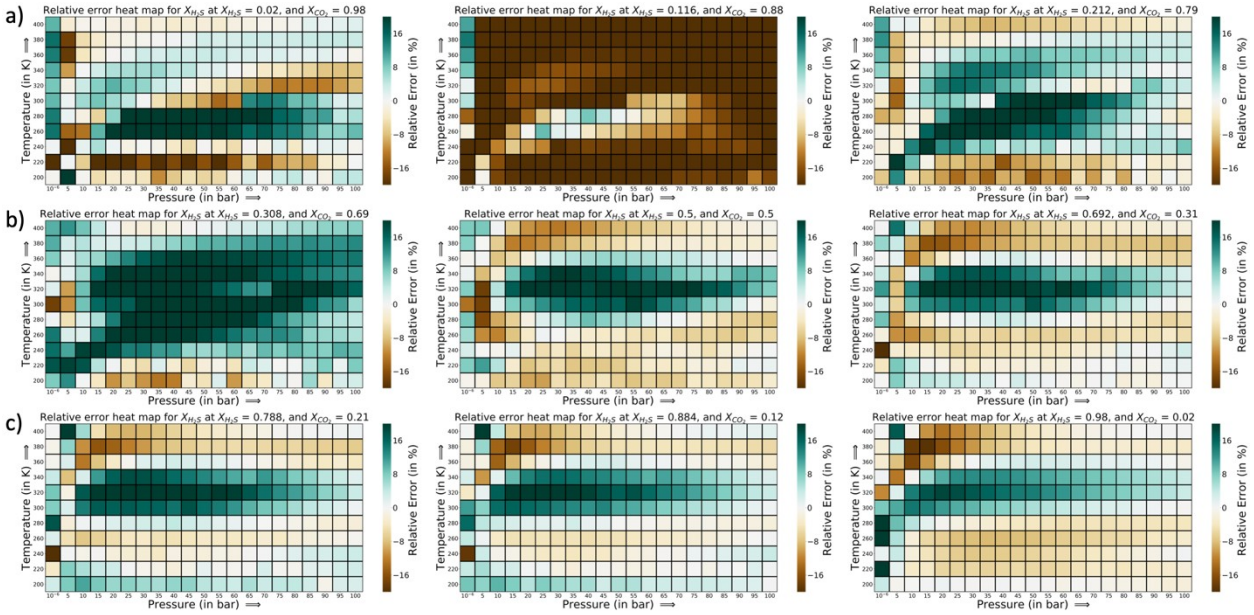

Figure S16. Relative error heat maps at the 90% PAC cut-off for H_2S in the H_2S-CO_2 mixture with triple-RBF kernel, a) X_{H_2S} = 0.02, 0. 116, and 0.212, b) X_{H_2S} = 0.308, 0.5, and 0.692, and c) X_{H_2S} = 0.788, 0.884, and 0.98.

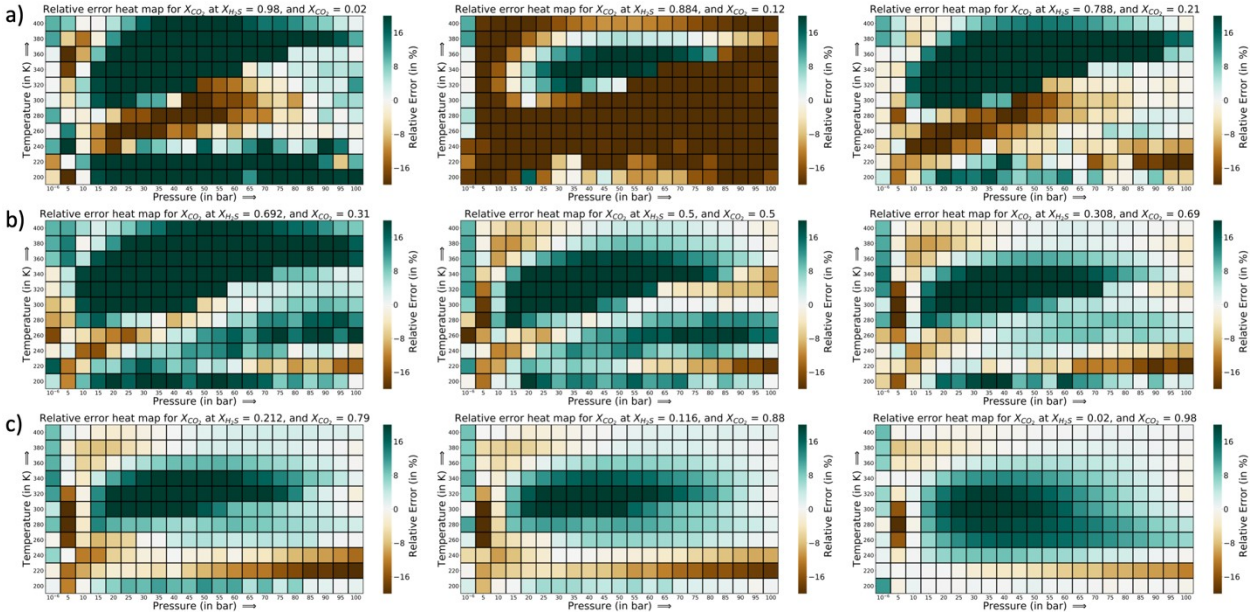

Figure S17. Relative error heat maps at the 90% PAC cut-off for CO_2 in the H_2S-CO_2 mixture with triple-RBF kernel, a) X_{CO_2}= 0.02, 0. 116, and 0.212, b) X_{CO_2}= 0.308, 0.5, and 0.692, and c) X_{CO_2}= 0.788, 0.884, and 0.98.

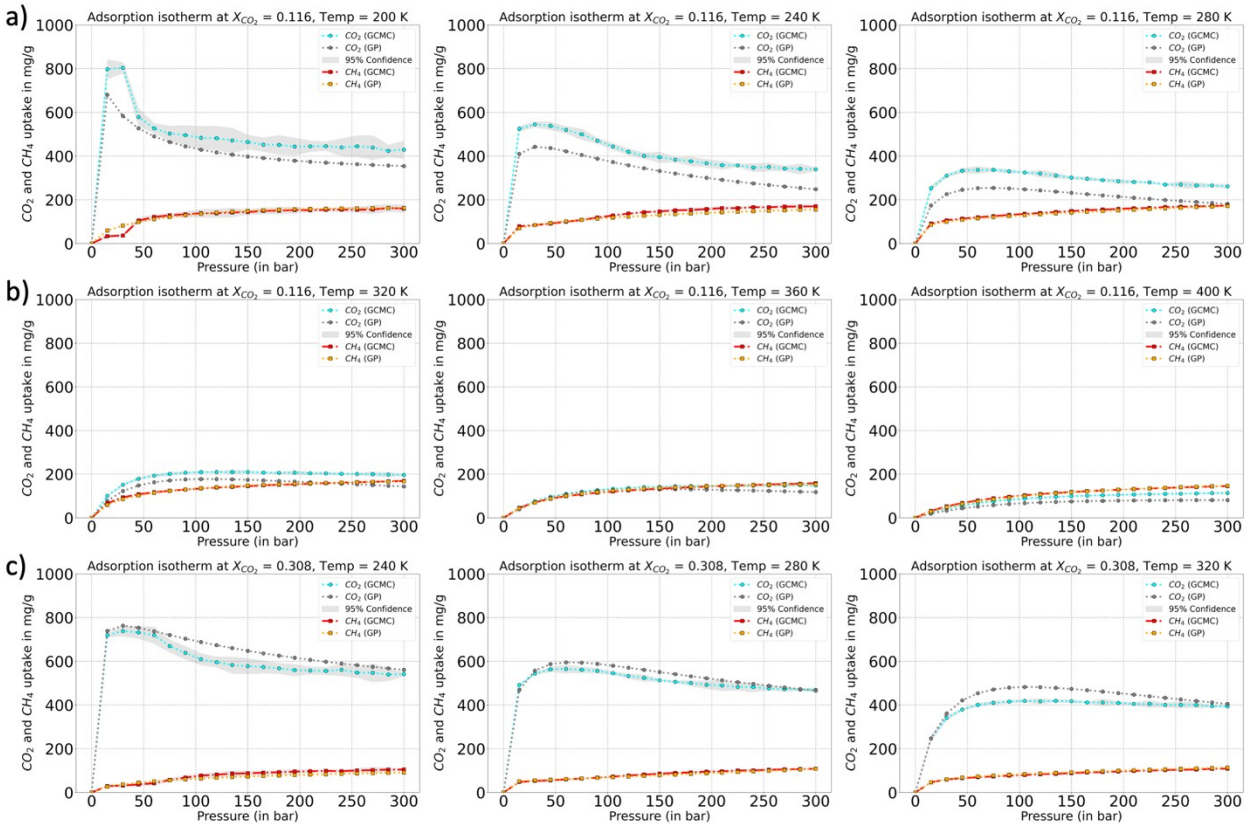

Figure S18. Adsorption isotherms at the 90% PAC cut-off for regions with highest relative errors for CO_2 in CO_2-CH_4 mixture (triple-RBF kernels), a) X_{CO_2} = 0.116 and T = 200, 240 and 280 K, b) X_{CO_2} = 0.116 and T = 320, 360 and 400 K, and c) X_{CO_2} = 0.308, 240, 280, and 320 K. We find the region at X_{CO_2} = 0.116 having the highest errors for CO_2 uptake. The model is also under-predicting adsorption of CO_2 at X_{CO_2} = 0.116, and then over-predicts from on X_{CO_2} = 0.308.

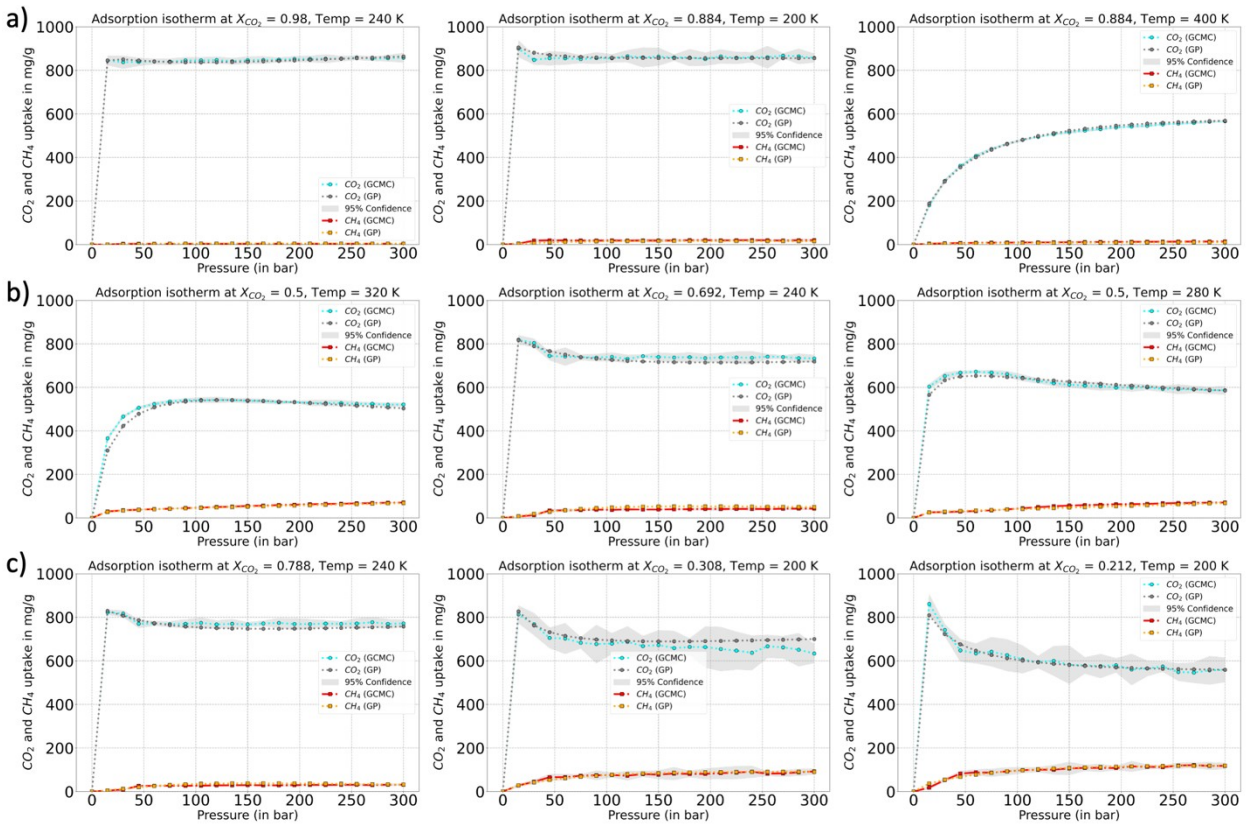

Figure S19. Adsorption isotherms at the 90% PAC cut-off for regions with highest relative errors for CH_4 in the CO_2-CH_4 mixture, a) $^{X_{CH_4}}$ = 0.02 ($^{X_{CO_2}}$= 0.98) T = 240 K, $^{X_{CH_4}}$ = 0.116 and T = 200 and 400 K, b) X_{CH_4}= 0.212 and T = 240 K, $^{X_{CH_4}}$= 0.308 and T = 240 K, $^{X_{CH_4}}$= 0.5 and T = 280 K, and c) $^{X_{CH_4}}$= 0.5 and T = 320 K, $^{X_{CH_4}}$= 0.692 and T = 200 K, $^{X_{CH_4}}$= 0.788 and T = 200 K. The region with most errors for CH_4 lies where adsorption is not very high. In fact, for a) it is evident that CH_4 uptake is near 0. Hence, the MRE is high at these points. The GP fit, we observe, corresponds in the same trend as the GCMC for all these regions.

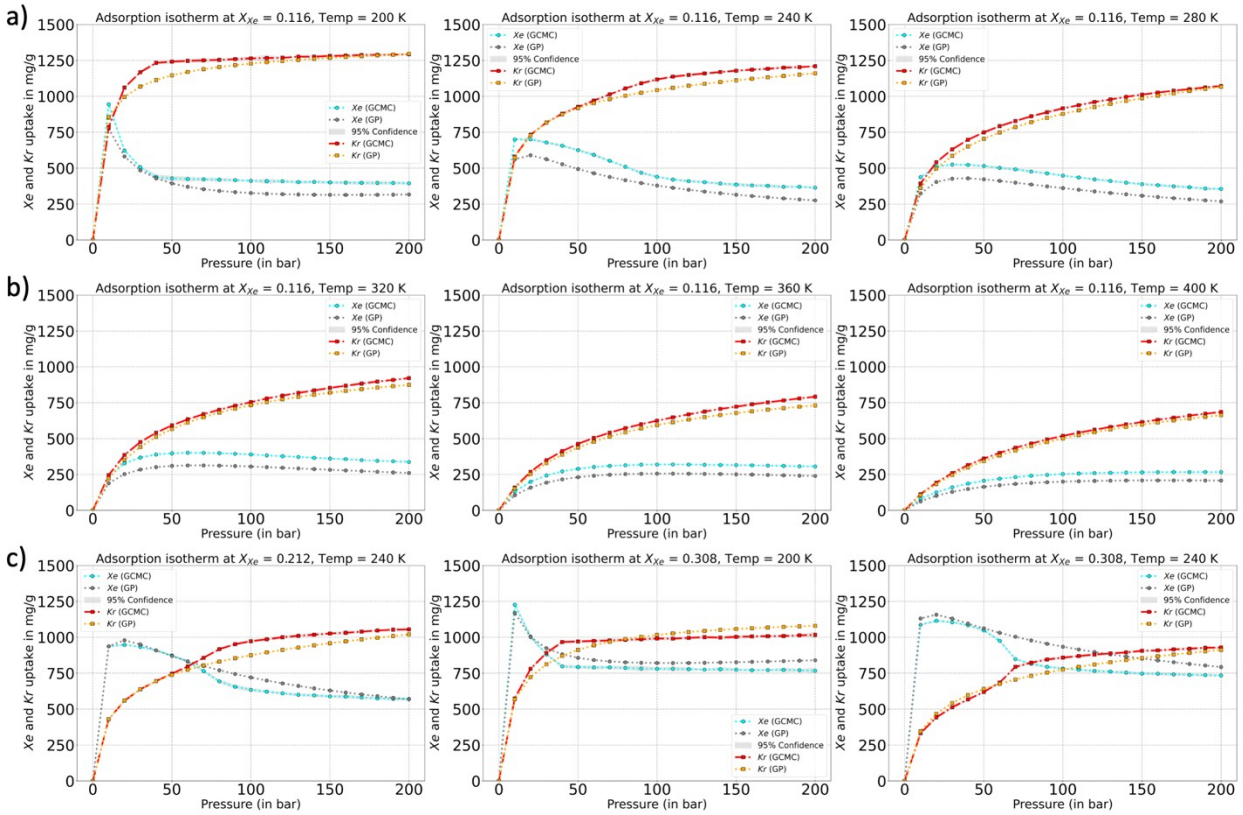

Figure S20. Adsorption isotherms at the 90% PAC cut-off for regions with highest relative errors for Xe in Xe-Kr mixture (triple-RBF kernels), a) X_{Xe} = 0.116 and T = 200, 240 and 280 K, b) X_{Xe} = 0.116 and T = 320, 360 and 400 K, and c) X_{Xe} = 0.212, 240 K, and X_{Xe} = 0.308 and T = 200, and 240 K. We find the region at X_{Xe} = 0.116 having the highest errors for Xe uptake. The model is also under-predicting adsorption of Xe at X_{Xe} = 0.116, and then slightly over-predicts from on X_{Xe} = 0.212 and 0.308.

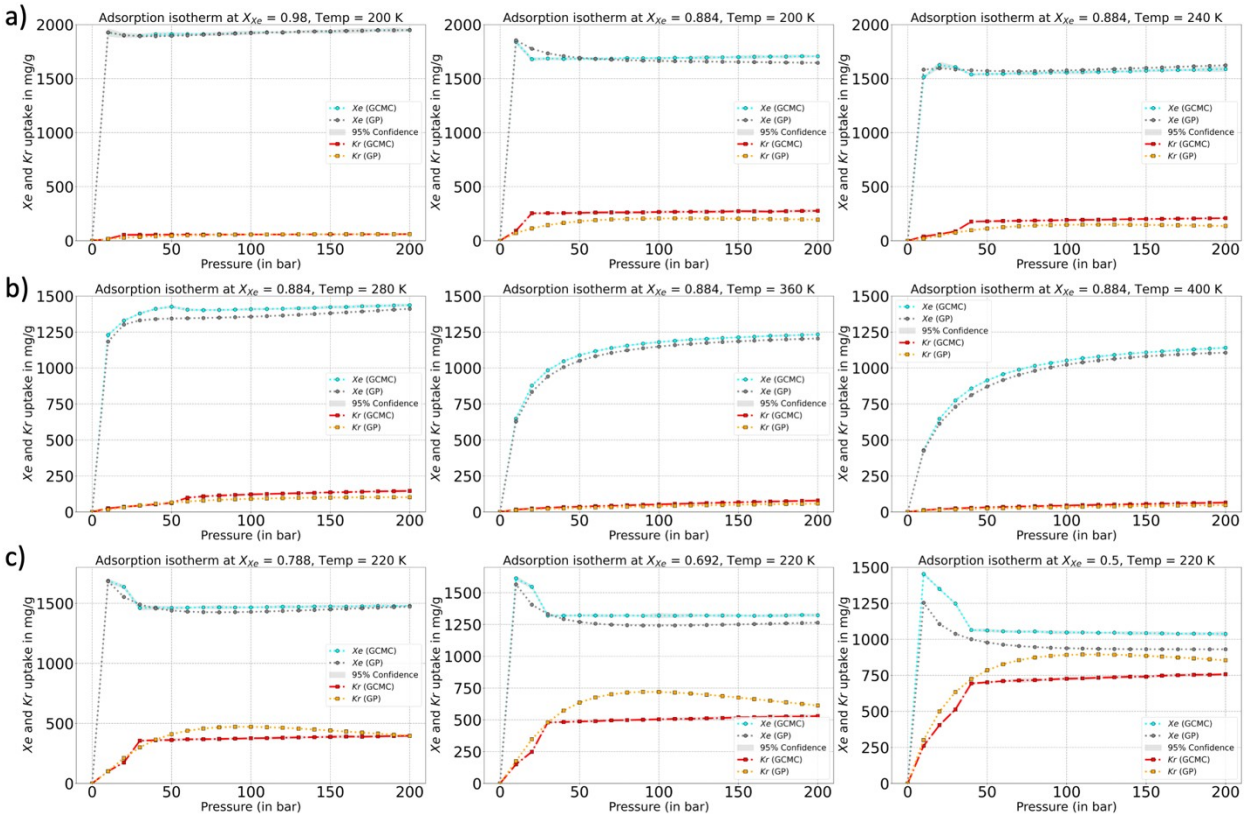

Figure S21. Adsorption isotherms at the 90% PAC cut-off for regions with highest relative errors for Kr in Xe-Kr mixture (triple-RBF kernels), a) X_{Kr} = 0.02 (X_{Xe} = 0.98), T = 200 K, and X_{Kr} = 0.116, and T = 200, and 280 K, b) X_{Kr} = 0.116 and T = 280, 360 and 400 K, and c) X_{Kr} = 0.212, 0.308, 0.50 at 220 K. We find the region at X_{Xe} = 0.116 having the highest errors for Xe uptake. The model is also under-predicting adsorption of Xe at X_{Xe} = 0.116, and then slightly over-predicts from on X_{Xe} = 0.212 and 0.308.

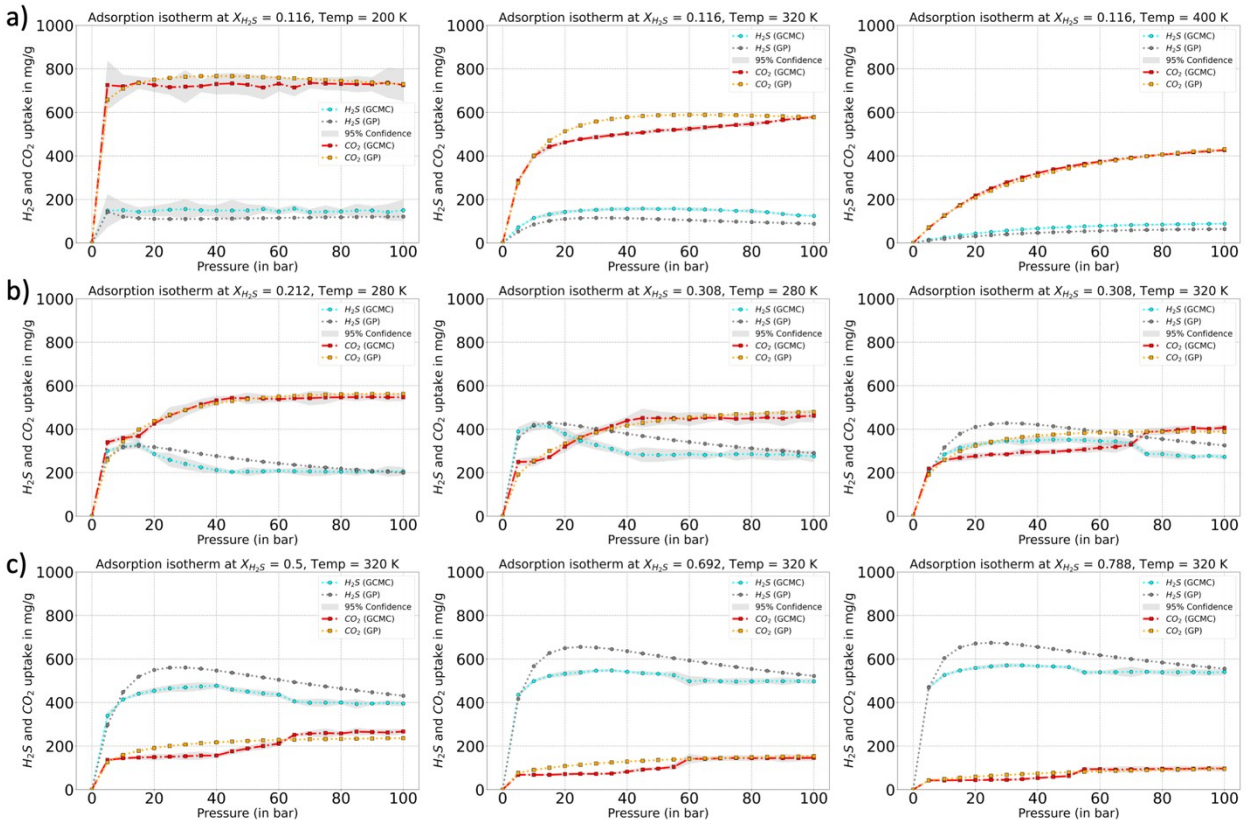

Figure S22. Adsorption isotherms at the 90% PAC cut-off for regions with highest relative errors for H_2S in H_2S-CO_2 mixture (RQ kernel), a) X_{H_2S} = 0.116 and T = 200, 320 and 400 K, b) X_{H_2S} = 0.212 and T = 280, X_{H_2S} = 0.308, T= 280 and 320 K, and c) X_{H_2S} = 0.5, and T = 320 K, and X_{H_2S} = 0.692 and T = 320 K, X_{H_2S} = 0.788, and T = 320 K. One observation is this, since both H_2S and CO_2 have strong interaction with the Cu-BTC MOF, the H_2S doesn't necessarily replace CO2 at low concentration of H_2S compared to how fast CO_2 itself was able to replace CH_4 at low CO_2 concentrations. These means the sharp inverse adsorption trend that was observed for CO2 in the CO_2-CH_4 mixture at low CO_2 concentrations are not seen here initially at low H_2S levels. But as the H_2S concentration increases to level of 30% (mole fraction of 0.308), we start to see those inverse trends in adsorption. Therefore, the errors trends are also pushed back, much more uniformly, across different mole fraction range.

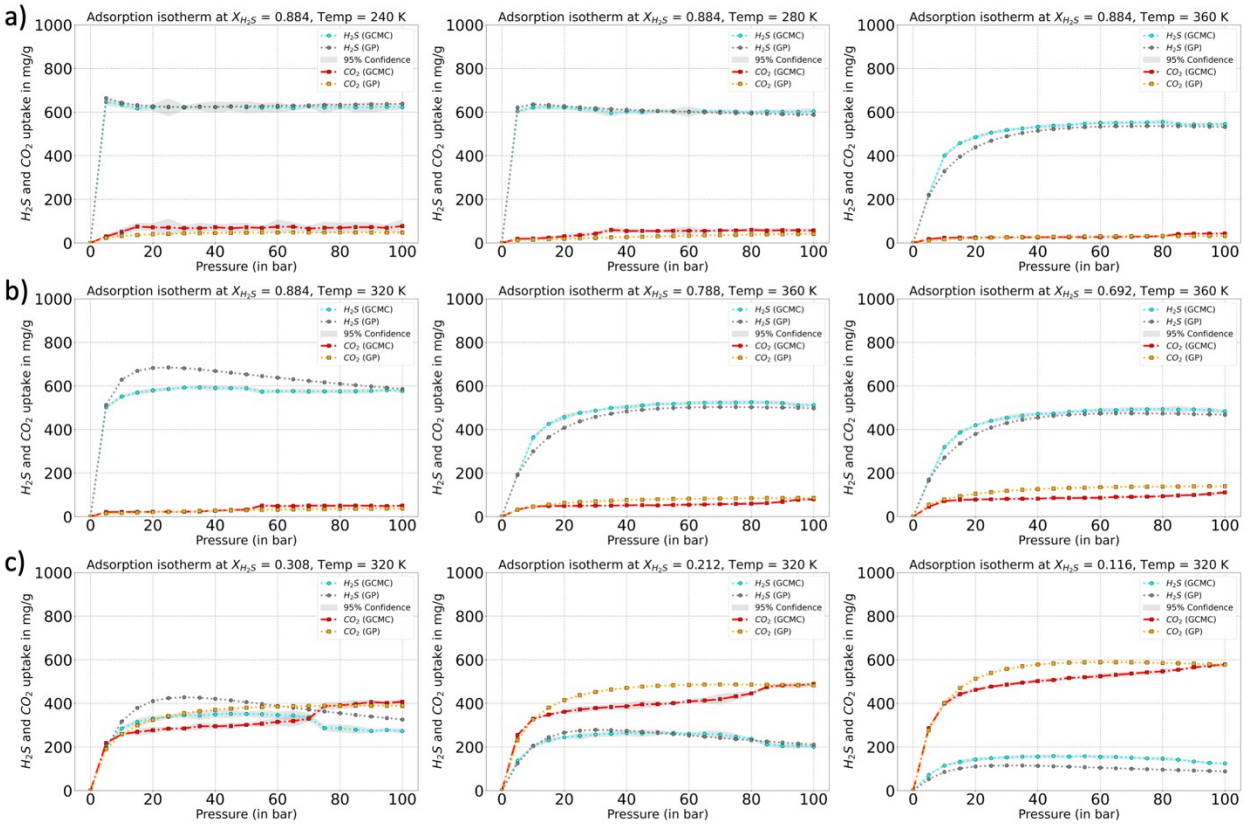

Figure S23. Adsorption isotherms at the 90% PAC cut-off for regions with highest relative errors for CO_2 in H_2S-CO_2 mixture (RQ kernel) for the P-X-T phase space, a) $X_{CO_2} = 0.116$ ($X_{H_2S} = 1 - X_{CO_2} = 0.884$) and T = 240, 280 and 360 K, b) $X_{CO_2} = 0.116$ and T = 320, $X_{CO_2} = 0.212$, and T= 360 K, $X_{CO_2} = 0.308$, and T= 360 K, and c) $X_{CO_2} = 0.692$, and T = 320 K, and $X_{CO_2} = 0.788$ and T = 320 K, $X_{CO_2} = 0.884$, and T = 320 K. We find the region at $X_{CO_2} = 0.116$ having the same trend as GCMC but under-prediction at low temperatures (T = 240 and 280 K). The error for CO_2 beyond this mole-fraction value is always over-prediction but again the trend is being followed.

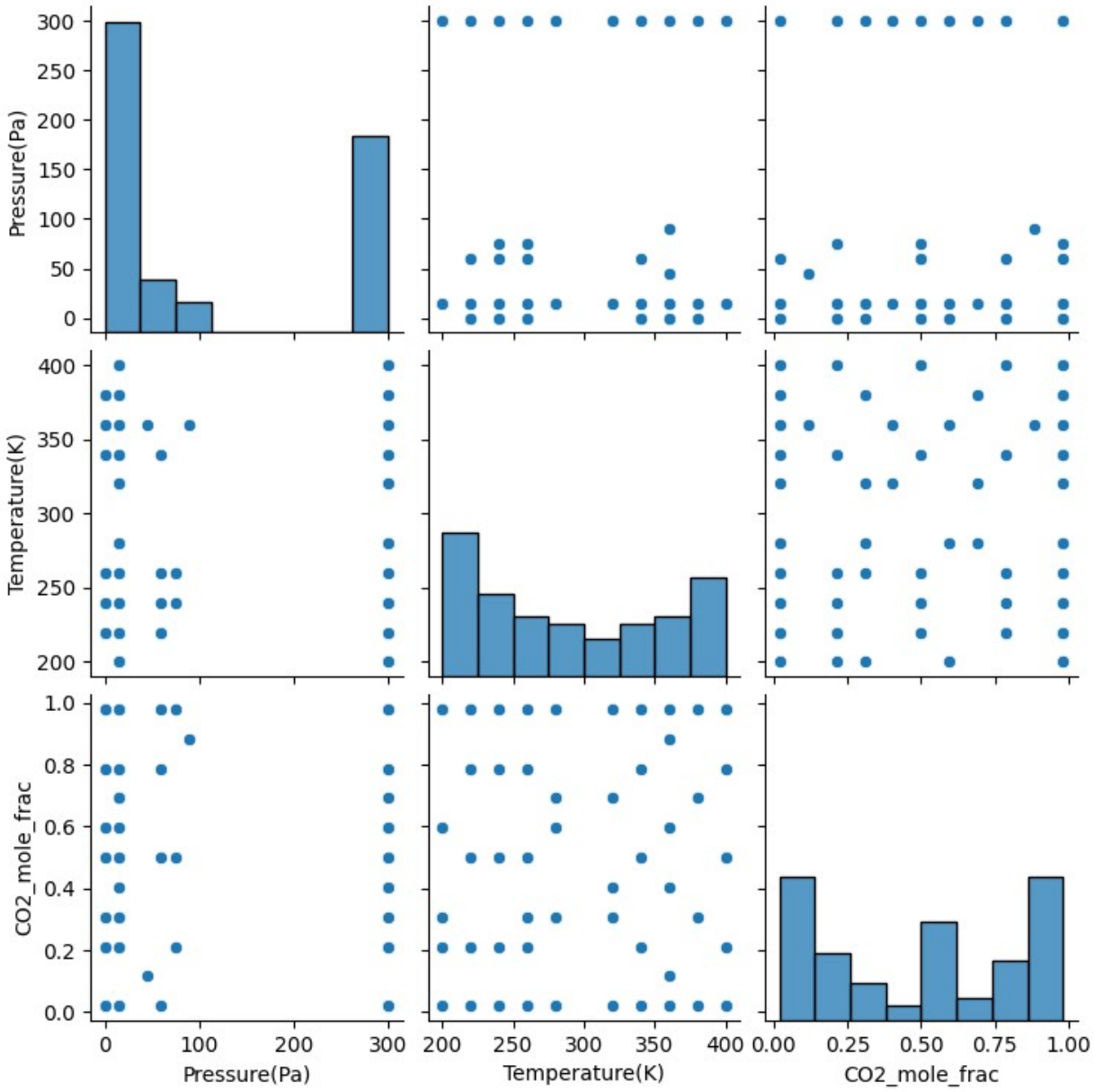

Figure S24. Correlation plots for CO_2-CH_4 mixture representing all the points sampled in the P-X-T after the AL is complete after 90% PAC limit is met. We see for pressure most points are located at low and high point values, followed by a similar distribution in the mole-fraction features. Finally, temperature has the most distribution with respect to the points sampled.

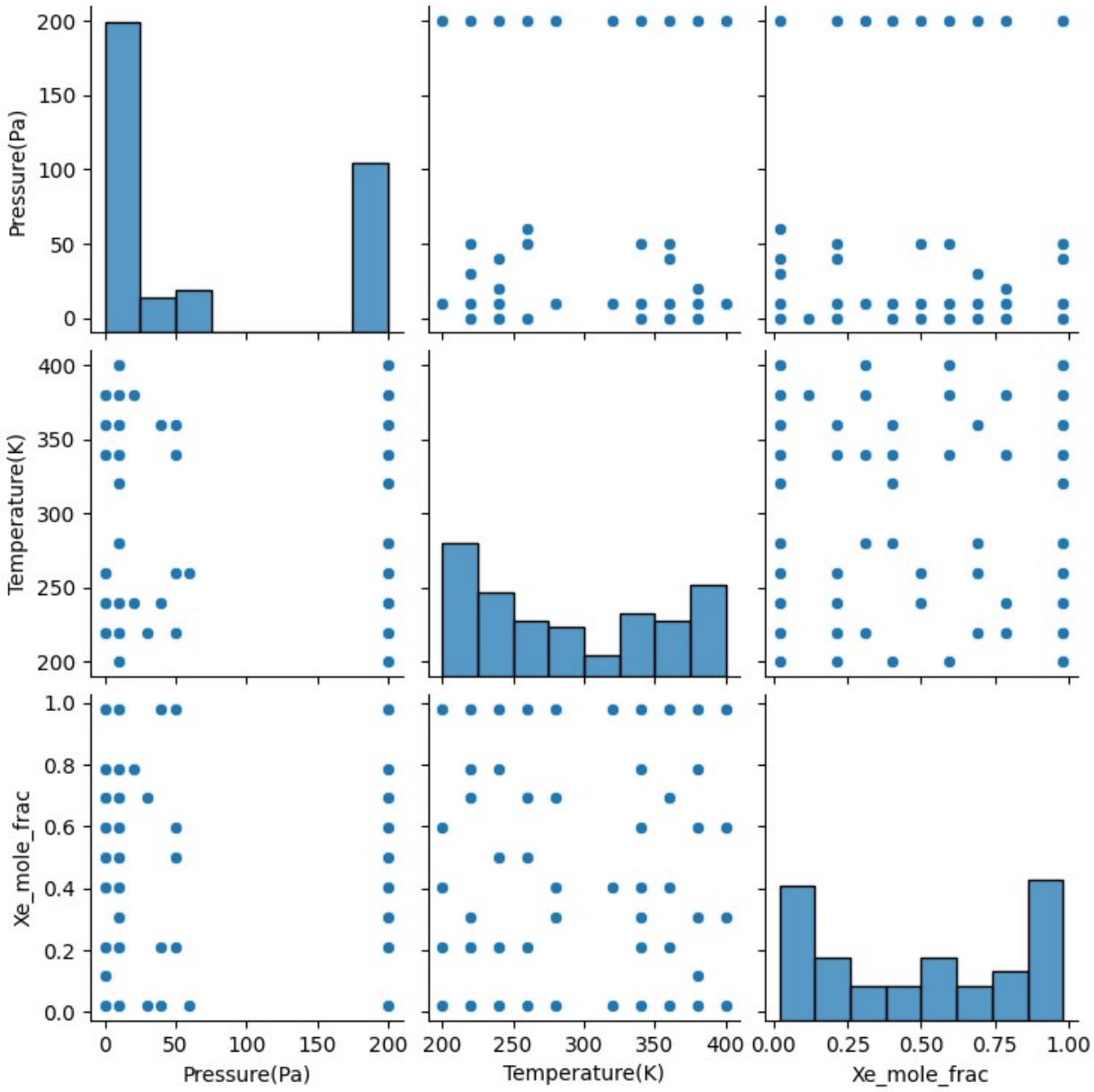

Figure S25. Correlation plots for Xe-Kr mixture representing all the points sampled in the P-X-T after the AL is complete after 90% PAC limit is met. We see for pressure most points are located at low and high point values, followed by a similar distribution in the mole-fraction features. Temperature again has the most distribution with respect to the points sampled.

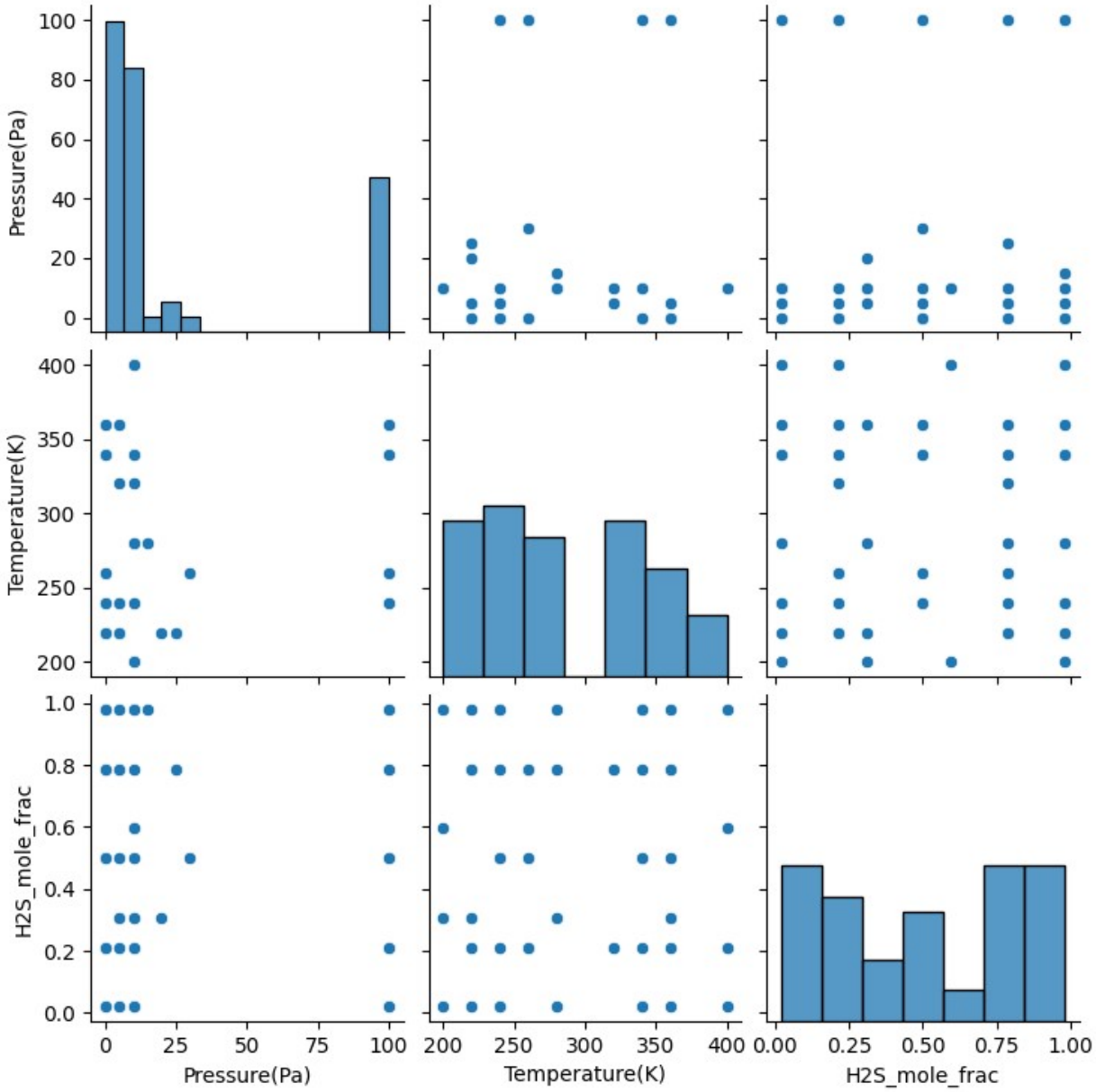

Figure S26. Correlation plots for H_2S-CO_2 mixture representing all the points sampled in the P-X-T after the AL is complete after 90% PAC limit is met. We see for pressure most points are located at low and high point values, followed by a similar distribution in the mole-fraction features. Here as well, temperature has the most distribution with respect to the points sampled.

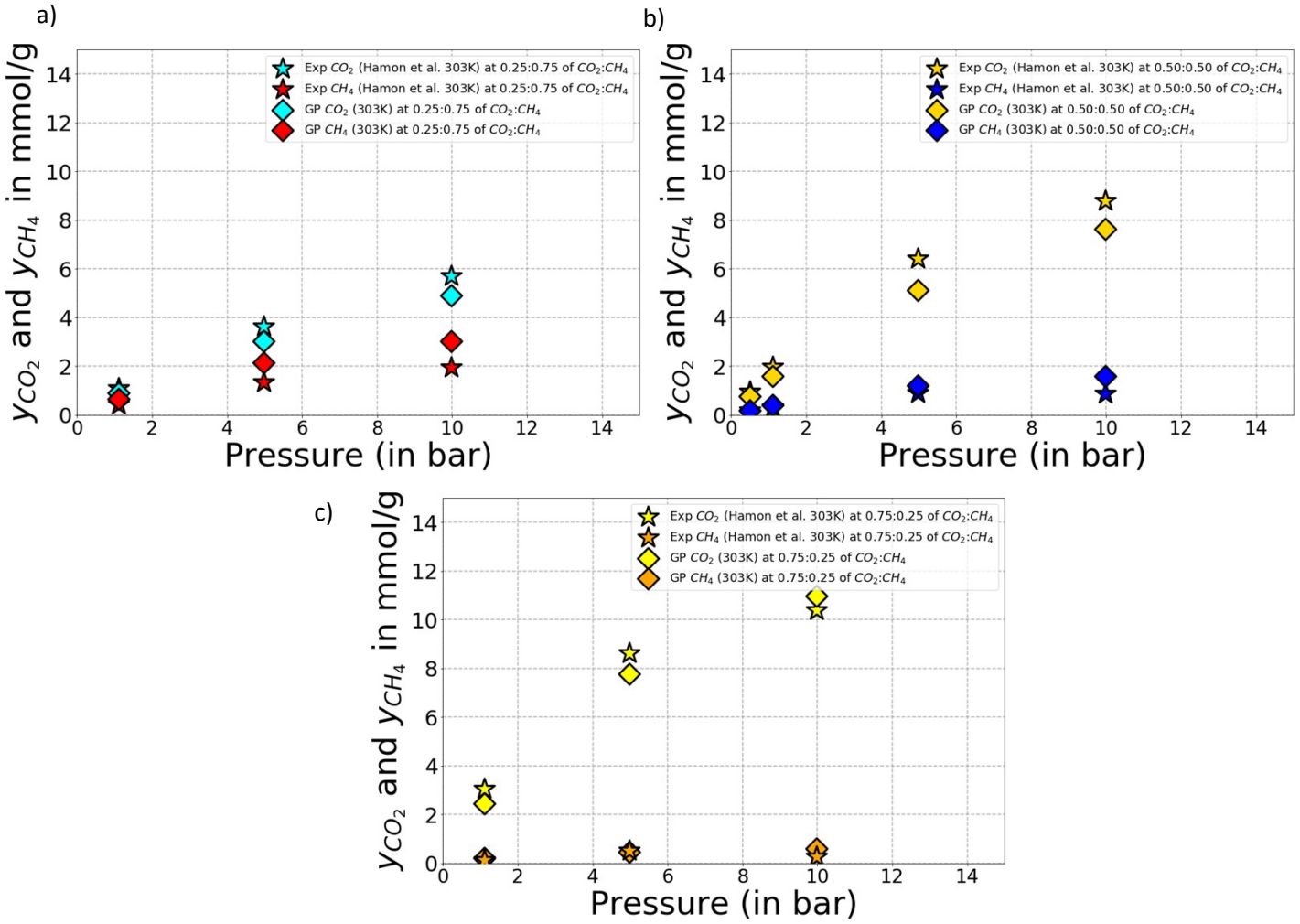

Figure S27. Comparing GP predicted CO_2 and CH_4 uptakes with experimental values for different ratios of CO_2 and CH_4 at 303K. Plot a) 0.25: 0.75 of CO_2:CH_4, b) 0.50: 0.50 of CO_2:CH_4, and c) 0.75: 0.25 of CO_2:CH_4. The GPs here were trained using AL for P-X-T for the CO_2-CH_4 mixture and it was terminated at PAC cut-off of 90%. We find very close agreement with GP predictions and experimental data. It is to be noted that the GPs were trained on GCMC data. Also, the experimental data was taken from Hamon et. al and the BISON dataset by Cai and coworkers was useful to obtain the data points. [1, 2]

References

1. L. Hamon, E. Jolimaître and G. D. Pirngruber, Industrial & Engineering Chemistry Research, 2010, 49, 7497–7503.

2. X. Cai, F. Gharagheizi, L. W. Bingel, D. Shade, K. S. Walton, and D. S. Sholl, Industrial & Engineering Chemistry Research, 2021, 60, 639–651.