Interpretable Learning of Complex Multicomponent Adsorption Equilibria from Self-attention

Yangzesheng Sun
University of Minnesota
Minneapolis, MN 55455
sun00032@umn.edu

Tyler R. Josephson
University of Minnesota
Minneapolis, MN 55455
josep180@umn.edu

J. Ilja Siepmann
University of Minnesota
Minneapolis, MN 55455
siepmann@umn.edu

Abstract

Modeling complex interactions in multicomponent adsorption equilibria is pivotal for the development of efficient chemical separation processes using nanoporous materials. Starting from the ideal-mixture adsorption model in statistical thermodynamics, we develop a machine learning formulation of multicomponent adsorption, Adsorptive Transformer, based on the self-attention mechanism. The Adsorptive Transformer accounts for the effects of guest-guest interactions on loading, and a "shallow" model is already able to significantly outperform a deep neural network on extrapolation in the state space. Our model is also readily interpretable to show adsorption mechanisms, such as cooperative adsorption and sieving.

1 Introduction

Measuring, modeling, and understanding multicomponent chemical systems with diverse types of interactions constitute one of the grand challenges in molecular and materials discovery, exemplified by the design of efficient and environmental-friendly chemical separation processes [1]. Although binary (2-component) and ternary (3-component) mixtures have been extensively studied [2, 3], real-world chemical systems may involve a larger number of interacting components, giving rise to difficulties in experimental measurements and predictive modeling. With an increased number of components, sampling the state space for equilibrium properties of the system becomes exponentially more difficult due to the well-known "curse of dimensionality".

Adsorption equilibria in nanoporous materials, such as zeolites and metal-organic frameworks, are of great interest for chemical separation [4]. The adsorption equilibrium $q(n_1, \cdots, n_k) = f(x_1, \cdots, x_{k-1}, p, T)$ gives the relationship between the amount of k chemical components contained in a porous material $(n_1, \cdots, n_k)$ and the thermodynamic state of the reservoir [5], generally mole fraction x_i, pressure p, and temperature T (for gas-phase adsorption, partial pressures $p_i = x_i p$ are typically used). Through highly selective adsorption or diffusion, nanoporous materials can provide more energy-efficient alternatives to conventional distillation-based separation [6, 7].

Herein, we study the adsorption equilibria for a series of 8-component systems of BTEX (benzene, toluene, ethylbenzene, o-xylene, m-xylene, and p-xylene) species mixed with hydrogen and ethane in an MFI-type zeolite structure. Separation of BTEX mixtures is among the most energy-consuming tasks in the chemical industry [8, 9], while MFI-type zeolite materials have been demonstrated as excellent candidates for efficient membrane separation of xylene isomers [10]. The molecular structures of the BTEX–ethane–hydrogen mixture, the pore structure of the MFI zeolite, and a snapshot of their adsorption are shown in Figure 1.

2 Related Work

Machine learning for nanoporous materials discovery Because of the vastness in the topological and chemical spaces of sorbent materials and guest molecules, machine learning methods have been

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io

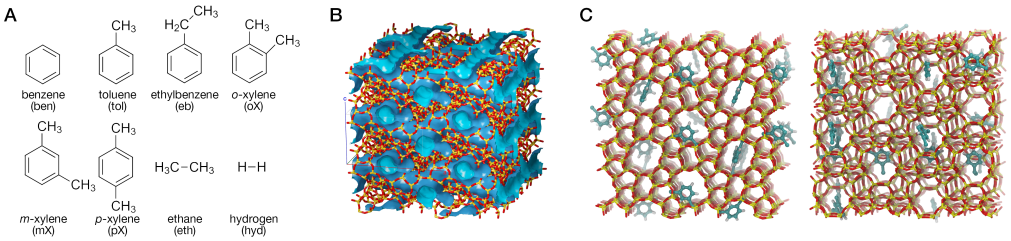

Figure 1: (A): Chemical constituents of the BTEX–ethane–hydrogen mixture. (B): The channel system of a $2 \times 2 \times 3$ supercell of the MFI zeolite with the pore surface highlighted in cyan. (C): Simulation snapshots of the mixture loading from two viewpoints.

widely employed for materials discovery [11, 12], and generative models have also been developed to propose new material structures [13, 14].

Learning interactions and relations in physical systems Physics-informed neural networks have been designed to model the interactions or relations between components in a system [15]. One of the most notable examples is the Behler–Parrinello network that implicitly models atomic interactions using symmetry functions [16]. By explicitly considering pairwise relations in a system, interactions can be also learned from the model, such as in Relation Networks [17] and Interaction Networks [18].

Physical interpretations of neural networks Recently, there has been growing interest in interpreting neural network architectures as physical systems [19]. The hallmark in this area is the Neural Ordinary Differential Equations that models an infinite-layer ResNet as an ODE system [20]. From a similar perspective, transformers can also be formulated as particle systems where the dynamics is governed by diffusion through inter-particle interactions and convection due to an external force [21].

3 Methods

Here we show that the multi-head self-attention mechanism, which is the core architecture of the transformer model widely used in natural language processing [22], can also constitute a physics-informed model for multicomponent adsorption equilibria. The simplest model for mixture adsorption equilibria can be obtained from statistical thermodynamics assuming no interaction between adsorbed molecules that can access the same set of adsorption sites [23]:

$$n_i = N \frac{\zeta_i(T) \mathrm{e}^{\beta \mu_i(p_i,T)}}{1 + \sum_j \zeta_j(T) \mathrm{e}^{\beta \mu_j(p_j,T)}}$$

where n_i is the (ensemble average) number of the ith type of molecule adsorbed, ζ_i is its molecular partition function, μ_i is its chemical potential, N is the total number of adsorption sites, and $\beta = 1/k_\mathrm{B}T$. It can be readily observed that by incorporating a fictitious "vacancy" component with partition function $\zeta_0 = 1$ and chemical potential $\mu_0 = 0$, the adsorption model becomes the celebrated softmax function,

$$n_i = N \frac{\mathrm{e}^{w_i(p_i,T)}}{\sum_j \mathrm{e}^{w_j(p_j,T)}}$$

where $w_i = \log \zeta_i(T) + \beta \mu_i(p_i, T)$ is a function of temperature and partial pressure of the ith component. Due to the variety in molecular structures and pore size distributions, different components usually exhibit different maximum adsorption capacities, and the strengths of the competitive adsorption (denominator of softmax) may not be uniform among all pairs of molecules. This yields

$$n_i = N_i \frac{\mathrm{e}^{w_{ii}(p_i,T)}}{\sum_j \mathrm{e}^{w_{ij}(p_i,p_j,T)}}$$

where n_i is now the diagonal of the dimension-wise softmax result on a "weight matrix" $\mathbf{W} = \{w_{ij}\}$. However, the influence of other components on the adsorption loading of one component is still limited to competitive adsorption by molecules displacing each other spatially instead of intermolecular interactions. The interaction between adsorbed molecules is reflected in the phenomenon of

cooperative adsorption [24], where the increase in one adsorbed component also leads to the increase in another attractively interacting component. Therefore, additional terms are introduced whose numerators contain the contributions from other components than the ith one, resulting in the inclusion of off-diagonal terms in the softmax matrix,

$$n_i = \sum_k N_{ik} \frac{\mathrm{e}^{w_{ik}(p_i, p_k, T)}}{\sum_j \mathrm{e}^{w_{ij}(p_i, p_j, T)}}$$

When $N_{ik} > 0$, the increase in partial pressure of kth component will lead to the increase in the adsorption of ith component, thus modeling cooperative adsorption. The softmax weight w_{ij} can be regarded as a metric describing how the partial thermodynamic states for two components "align" with each other in producing interactions, hence it is naturally represented as a dot product [25]

$$w_{ij}(p_i, p_j, T) = \mathbf{q}_i(p_i, T)\mathbf{k}_j(p_j, T)'$$

where $\mathbf{q}$ and $\mathbf{k}$ are vector representations of the partial thermodynamic states of a component. The prime symbol denotes transpose to avoid confusion with the temperature T. Finally, our physics-informed model for multicomponent adsorption can be written in a matrix form as

$$\mathbf{n} = \mathrm{softmax}(\mathbf{QK}') \cdot \mathbf{N}$$

It is noteworthy that the model is almost identical to the self-attention mechanism $\mathbf{Z} = \mathrm{softmax}(\mathbf{QK}')\mathbf{V}$ (omitting the scaling factor for brevity) [22], with the only differences being prepending an additional column of zeros in $\mathbf{K}$ and using a row-wise dot product over the maximum adsorption matrix $\mathbf{N}$ instead of matrix multiplication. Same as the transformer, we use a direct linear projection for $\mathbf{Q}$ and $\mathbf{K}$ from the thermodynamic state vector (p_i, p, T) of each component. The physical interpretation of multi-head attention is also straightforward. Each attention head represents a separate adsorption domain in the adsorbent material, and the total adsorption is the sum of contributions from each domain. The existence of multiple adsorption domains in the structure of a nanoporous material has also been observed in experimental studies [26].

4 Results and Discussion

Here, we report our preliminary results of the physics-informed model for adsorption equilibria, named Adsorptive Transformer, on the adsorption of BTEX–ethane–hydrogen mixture in the MFI zeolite using high-throughput Monte Carlo simulations performed for 27,648 different combinations of mole fractions for each component, total pressure, and temperature (see Appendix for simulation details). To illustrate the intrinsic advantage of Adsorptive Transformer in predicting the state space surface of adsorption equilibria, the model was trained only on 6- or 7-component simulation data without hydrogen, ethane, or both, (62.5%) and was tested on the simulations with all 8 components (37.5%). Figure 2 shows the training and test results for the Adsorptive Transformer on the 8-component BTEX dataset compared with a multi-layer perceptron (MLP) neural network as baseline.

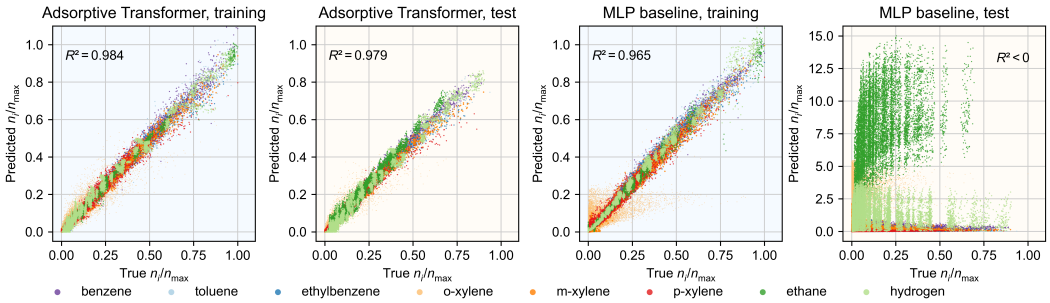

Figure 2: Predicting multicomponent adsorption equilibria using the Adsorptive Transformer and an MLP as baseline. The number of molecules for each component was normalized to the maximum number observed in the dataset. The size of the scatter points in each color was scaled by the maximum number of molecules in the corresponding type. R^2 values were calculated as the coefficient of determination.

Although the MLP attained a similar performance as the Adsorptive Transformer in the 6- or 7-component training set, it gave a completely erroneous prediction deviating from simulation results by

orders of magnitude on the 8-component test set, because extensive extrapolation is required to predict additional degrees of freedom in the state space. On the contrary, the Adsorptive Transformer gave a prediction as accurate as the training set when extending to 8 components, even if there is no training example where hydrogen and ethane coadsorb. By directly incorporating strong physical inductive biases about adsorption equilibria with the self-attention mechanism, we have achieved significantly improved generalization in the state space compared to a model without domain knowledge.

Since the attention query $\mathbf{Q}$ and key $\mathbf{K}$ are obtained from linear transformations and are independent to the type of adsorbed molecule in a system, maximum adsorption matrices $\mathbf{N}$ consist of the majority of parameters in the model. The learned weights of $\mathbf{N}$ are straightforwardly interpretable by visualizing these weight matrices for the trained Adsorptive Transformer (see Figure 3); here N_{ii} represents the maximum attainable adsorption (saturation capacity) of component i without presence of any other type of molecules, and N_{ij} represents the cooperative or obstructive effects of component j on component i. We apply L_1 regularization on the non-diagonal terms of $\mathbf{N}$ matrices to ensure that the total adsorption of a component is still mainly determined by itself. Among 8 components, three xylene isomers have the strongest steric hindrance, thus their locations in the MFI zeolite structure when adsorbed are highly confined (see Figure 1C). This is reflected in the weight matrices that only two of the four adsorption domains have large diagonal elements for these molecules. Benzene, toluene, and ethylbenzene are slightly less hindered, allowing for more possible adsorption locations and making their corresponding weights active in all four adsorption domains learned. Hydrogen and ethane, however, are much smaller molecules than aromatic compounds, thus their adsorption behavior and the distribution of weights differ significantly. The largely positive rows of weights on these molecules in one of the adsorption domains indicate strong cooperative adsorption effects, which is indeed confirmed by our simulation results that show their loading to be positively correlated with the total uptake of aromatic compounds.

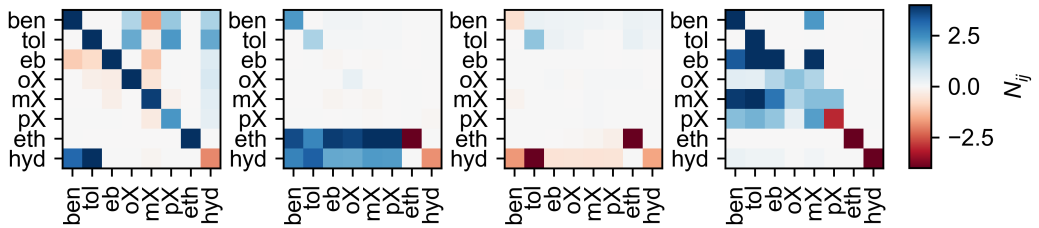

Figure 3: Weight matrices $\mathbf{N}$ in each attention head of the trained Adsorptive Transformer. Weights are clamped into $[-4, 4]$ for better visualization. Four attention heads are used since the average test set error does not further decrease when increasing the number of heads above 4.

5 Conclusions

While comprehensive measurements in high-dimensional state space for multicomponent chemical systems can be prohibitively expensive, our Adsorptive Transformer provides a new route toward solving this challenge as it is able to generalize from mixtures with fewer components to those with more components. Our preliminary results focus on "shallow" Adsorptive Transformers, deeper models can be constructed analogously as for the regular transformer by stacking multiple attention layers and pointwise feedforward layers. Our derivation for the Adsorptive Transformer also supplies a novel view of interpreting the attention mechanism as a thermodynamic equilibrium. We hope that this work will inspire research leading to the development of physics-informed machine learning models for chemical and materials discovery.

Acknowledgement

This research is primarily supported by the U.S. Department of Energy (DOE), Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences and Biosciences under Award DE-FG02-17ER16362. This research used resources of the Argonne Leadership Computing Facility, a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357. Additional computer resources were provided by the Minnesota Supercomputing Institute at the University of Minnesota. Y. S. is grateful for support from the Robert and Jill DeMaster-Excellence Fellowship.

References

[1] J. I. Siepmann, J. F. Brennecke, D. T. Allen, M. T. Klein, P. E. Savage, G. C. Schatz, and F. M. Winnik. ACS virtual issue on multicomponent systems: Absorption, adsorption, and diffusion. *J. Chem. Eng. Data*, 63:3651, 2018.

[2] H. Orbey and S.I. Sandler. *Modeling Vapor-Liquid Equilibria: Cubic Equations of State and Their Mixing Rules*. Cambridge University Press, 1998.

[3] J. M. Prausnitz, F. F. Anderson, and T. F. Anderson. *Computer Calculations for Multicomponent Vapor-liquid and Liquid-liquid Equilibria*. Prentice-Hall, 1980.

[4] S. Kitagawa. Future porous materials. *Acc. Chem. Res.*, 50:514–516, 2017.

[5] R.T. Yang. *Gas Separation By Adsorption Processes*. World Scientific Publishing Company, 1997.

[6] P. Bernardo, E. Drioli, and G. Golemme. Membrane gas separation: A review/state of the art. *Ind. Eng. Chem. Res.*, 48:4638–4663, 2009.

[7] J.-R. Li, R. J. Kuppler, and H.-C. Zhou. Selective gas adsorption and separation in metal–organic frameworks. *Chem. Soc. Rev.*, 38:1477–1504, 2009.

[8] D. S. Sholl and R. P. Lively. Seven chemical separations to change the world. *Nature*, 532:435–437, 2016.

[9] Y. Yang, P. Bai, and X. Guo. Separation of xylene isomers: A review of recent advances in materials. *Ind. Eng. Chem. Res.*, 56:14725–14753, 2017.

[10] M. Y. Jeon, D. Kim, P. Kumar, P. S. Lee, N. Rangnekar, P. Bai, M. Shete, B. Elyassi, H. S. Lee, K. Narasimharao, S. N. Basahel, S. Al-Thabaiti, W. Xu, H. J. Cho, E. O. Fetisov, R. Thyagarajan, R. F. DeJaco, W. Fan, K. A. Mkhoyan, J. I. Siepmann, and M. Tsapatsis. Ultra-selective high-flux membranes from directly synthesized zeolite nanosheets. *Nature*, 543:690–694, 2017.

[11] K. M. Jablonka, D. Ongari, S. M. Moosavi, and B. Smit. Big-data science in porous materials: Materials genomics and machine learning. *Chem. Rev.*, 120:8066–8129, 2020.

[12] H. Lyu, Z. Ji, S. Wuttke, and O. M. Yaghi. Digital reticular chemistry. *Chem*, 6:2219–2241, 2020.

[13] B. Kim, S. Lee, and J. Kim. Inverse design of porous materials using artificial neural networks. *Sci. Adv.*, 6, 2020.

[14] Z. Yao, B. Sanchez-Lengeling, N. S. Bobbitt, B. J. Bucior, S. G. H. Kumar, S. P. Collins, T. Burns, T. K. Woo, O. Farha, R. Q. Snurr, and A. Aspuru-Guzik. Inverse design of nanoporous crystalline reticular materials with deep generative models, 2020. ChemRxiv:12186681.

[15] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.*, 71:361–390, 2020.

[16] J. Behler and M. Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98:146401, 2007.

[17] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30*. 2017.

[18] P. Battaglia, R. Pascanu, M. Lai, D. Jimenez Rezende, and K. Kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Advances in Neural Information Processing Systems 29*, pages 4502–4510. 2016.

[19] L. Ruthotto and E. Haber. Deep neural networks motivated by partial differential equations. *Journal of Mathematical Imaging and Vision*, 62:352–364, 2020.

[20] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems 31*, pages 6571–6583. 2018.

[21] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T.-Y. Liu. Understanding and improving transformer from a multi-particle dynamic system point of view, 2019. arXiv:1906.02762. arXiv.org e-Print archive. https://arxiv.org/abs/1906.02762.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008. 2017.

[23] M. Jaroniec. Statistical thermodynamics of mixed-gas adsorption. Localized monolayer adsorption on heterogeneous surfaces. *J. Chem. Soc., Faraday Trans. 2*, 73:933–938, 1977.

[24] G. Li, P. Xiao, and P. Webley. Binary adsorption equilibrium of carbon dioxide and water vapor on activated alumina. *Langmuir*, 25:10666–10675, 2009.

[25] A. Graves, G. Wayne, and I. Danihelka. Neural turing machines, 2014. arXiv:1410.5401. arXiv.org e-Print archive. https://arxiv.org/abs/1410.5401.

[26] H. S. Cho, J. Yang, X. Gong, Y.-B. Zhang, K. Momma, B. M. Weckhuysen, H. Deng, J. K. Kang, O. M. Yaghi, and O. Terasaki. Isotherms of individual pores by gas adsorption crystallography. *Nat. Chem.*, 11:562–570, 2019.

[27] M. G. Martin and J. I. Siepmann. Novel configurational-bias monte carlo method for branched molecules. transferable potentials for phase equilibria. 2. united-atom description of branched alkanes. *J. Phys. Chem. B*, 103:4508–4517, 1999.

[28] A. Z. Panagiotopoulos, N. Quirke, M. Stapleton, and D. J. Tildesley. Phase equilibria by simulation in the gibbs ensemble - alternative derivation, generalization and application to mixture and membrane equilibria. *Mol. Phys.*, 63:527–545, 1988.

Appendix

Molecular Simulation Details Configurational-bias Monte Carlo simulations [27] are performed in the isothermal-isobaric version of the Gibbs ensemble [28]. In these simulations, N_i molecules for each of i species can distribute between two simulation boxes, a zeolite box and a vapor-phase reservoir. The total composition N_i^{total}, temperature T, and pressure p are specified as the independent variables. Monte Carlo moves are used to translate and rotate molecules within a given box, as well as to exchange molecules between boxes and change the volume of the vapor box. Each simulation is equilibrated for at least 20000 MC cycles, after which at least 40000 MC cycles are used for production. During production, the average number of molecules of each type adsorbed in the zeolite box (N_i^{zeo}) and remaining in the vapor box (N_i^{vapor}) is measured — these are the dependent variables.