TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Johnsson, E., Sharma, S., Rao, A. G., Dubbeldam, D., Calero, S., & Vlugt, T. J. H. (2026). Predicting the Maximum Loading in Zeolites for Hydroisomerization Applications: A Machine Learning Approach. *Journal of Physical Chemistry C*, *130*(11), 4299-4314. https://doi.org/10.1021/acs.jpcc.5c08611

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

*This work is downloaded from Delft University of Technology.*

THE JOURNAL OF PHYSICAL CHEMISTRY C
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCC

Article

# Predicting the Maximum Loading in Zeolites for Hydroisomerization Applications: A Machine Learning Approach

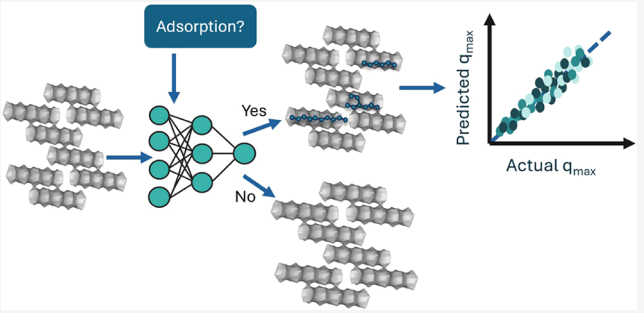
Eric Johnsson, Shrinjay Sharma, Arvind Gangoli Rao, David Dubbeldam, Sofia Calero, and Thijs J. H. Vlugt*

Cite This: J. Phys. Chem. C 2026, 130, 4299–4314

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Hydroisomerization of alkane isomers is an important step in the manufacture of current kerosene and sustainable aviation fuels. Zeolites are used as acid catalysts in this process. It is therefore important to have predictions of the adsorption capacity or maximum loading of hydrocarbons in zeolites. Here, a cascade model using machine learning models is used to predict the maximum loading of alkane isomers in zeolites. The cascade is composed of a gradient-boosted tree classifier stage that predicts whether adsorption occurs and a regressor predicting the value of the maximum loading. The final data set consists of 45 different adsorbates (both linear and branched alkanes up to $C_{16}$) and 97 different zeolite structures, resulting in 4365 data points. Descriptors include information on the geometry and topology of zeolite channels as well as the shape and size of the adsorbates. Extra composite descriptors are also present to provide the physical basis for predictions. Multiple regressors of different natures are considered: support vector regressors, gradient-boosted trees, extreme gradient-boosted trees, and the TabPFN pretrained model. TabPFN yields the highest generalization performance and the lowest error. An interpretability analysis using SHAP reveals that the most influential descriptors are physically meaningful, highlighting steric and volumetric constraints as the primary factors controlling the prediction of $q_{max}$. It is shown that despite both the classifier and the regressor being insensitive to random splits in data, the regressor is prone to overfitting at low fractions of data withheld for testing. The cascade model is compared to an Artificial Neural Network for training and resource efficiency. Despite training being longer for the neural network, the final model is lighter in both memory and storage. This work is built on our previous research in predicting the Henry coefficients of long-chain alkanes in zeolites. Using this previous model and the findings of this work, one could construct the adsorption isotherm for any alkane, thus enabling the analysis of adsorption behavior of alkane mixtures using IAST.



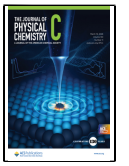

Downloaded via TU DELFT on April 2, 2026 at 09:04:49 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

## 1. INTRODUCTION

Climate change has been one of the most pressing issues in recent years, being the subject of international agreements.[1] This is pushing fossil-fuel-dependent industries to research more sustainable practices. For the aviation industry (which accounts for about 2.5% of global carbon emissions[2]), solutions to reduce the industry's footprint without sacrificing passenger demand are to either redesign aircraft components or try and switch to a more environmentally friendly fuel.[3] One such fuel is Sustainable Aviation Fuel (SAF)[3,4] produced from nonfossil or low-carbon feedstocks such as lipids[5] (e.g., used cooking oil or animal fats), biomass-derived fuels,[3] or syngas routes that qualify as SAF when combined with carbon capture utilization and storage and when meeting lifecycle greenhouse gas reduction and sustainability criteria (e.g., CORSIA[6] and EU RED[7]). SAF promises to significantly reduce carbon emissions while still being compatible with current engines and fuel infrastructure (both for processing and distribution).[3]

The most commercially viable way to make SAF is through the HEFA pathway (Hydroprocessed Esters and Fatty Acids),[3] in which natural oils and fats are transformed into usable jet fuel. This pathway uses a 2-fold process of deoxygenation and hydroisomerization.[3] The latter, also known as catalytic dewaxing,[8] transforms linear alkane chains into branched isomers by means of a bifunctional catalyst composed of a metal (generally platinum) and a zeolite. Linear alkanes are first delivered to the metal site, where their transformation into olefins occurs. These olefins then diffuse to the zeolite, where Brønsted acid sites are encountered, changing these olefins to carbenium ions. Subsequent deprotonation is carried out,

Received: December 19, 2025
Revised: March 3, 2026
Accepted: March 4, 2026
Published: March 10, 2026



ACS Publications
© 2026 American Chemical Society



https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299–4314

resulting in branched alkenes. By a final step of hydrogenation, the alkenes are transformed into branched alkanes.[9] Making alkanes undergo hydroisomerization offers multiple benefits when looking at fuel performance. First, it improves cold-flow properties (e.g., viscosity and freezing point).[10] Second, it increases the octane rating of the fuel, making it less prone to ignite under compression.[10]

Despite the current understanding of hydroisomerization, important knowledge gaps such as competitive adsorption or optimal zeolite topology for a given process still need to be addressed.[11] Carrying out experiments that can provide sufficient information at the molecular level is difficult to set up. Experimental adsorption studies of long-chain alkanes typically focus on single-component systems; consequently, the effects of competitive adsorption and desorption are frequently not accounted for.[12] As an alternative to adsorption experiments, molecular simulations using Monte Carlo simulations in the grand-canonical ensemble can be performed.[13] Despite these advancements, a larger problem has yet to be addressed: the large combinatorial space between alkanes and zeolites. The International Zeolite Association (IZA) recognizes 261 synthesizable zeolite structures to date, and the number goes to millions when considering hypothetical structures.[14−16] The number of possible isomers also explodes for long-chain alkanes.[17] This renders current methods such as molecular simulation very inefficient for screening purposes due to large computational requirements. There is thus a need for a fast, cheap, and reliable method to predict adsorption properties of alkane isomers in zeolites. Recently, an effort has been made to predict the Henry coefficient $K_H$ that describes adsorption at low pressures.[18] If one assumes a single-site Langmuir-type behavior, an expression for the full isotherm can be made with both the Henry coefficient and the maximum loading.[19] In more complex systems exhibiting multistep adsorption, the intermediate-pressure regime may deviate from single-site Langmuir behavior.[20,21] Nevertheless, $K_H$ still captures the correct zero-loading limit, and $q_{max}$ defines the maximum loading. Even if inflection points in isotherms are not reproduced, the two parameters provide a well-defined thermodynamic envelope, binding the isotherm between its low- and high-pressure limits. The isotherm also serves as input for Ideal Adsorbed Solution Theory (IAST)[22−24] for studying the adsorption properties of alkane mixtures.[25] Therefore, there is a need to have a way to compute the maximum loading of long-chain alkanes in zeolites.

To address the problem of large combinatorial space, Machine Learning (ML) algorithms have become increasingly popular due to their data-driven approach. This allows one to capture complex relations between a set of inputs (called descriptors) and outputs (called targets), all while being computationally less expensive than classical molecular simulations or experiments. In recent years, ML algorithms have been extensively used in research on adsorption in microporous materials.[26−28] Some of the most prevalent uses are as tools for High-Throughput Computational Screening (HTCS) or Quantitive Structure−Property Relationships (QSPR). As highlighted in the reviews by Altintas et al.[29] and Yang et al.[28] on the use of ML with Metal−Organic Frameworks (MOFs), ML is capable of using data pertaining to pore size and geometry to screen structures for the best candidate if given a specific property. Xue et al.[30] used a random forest model to find the most viable all-silica zeolites for the separation of propane and propylene. An example of ML used for QSPR includes the analysis by Xiuying et al.[31] to find the governing structural parameters behind the adsorption selectivity of $CO_2$ and $N_2$ in all-silica zeolites. Tatlier et al.[32] reported on the relation between the water uptake in zeolites and the structural and chemical properties of the framework. ML models can also be used to directly predict a large set of diverse properties. Evans and Coudert[33] used extreme gradient-boosted trees to determine mechanical properties such as the shear and bulk moduli of zeolites, providing insight into which quantities impact these properties the most. Another work is the one of Yu et al.,[34] that predicts the Henry coefficient of several small molecules such as hydrogen and $CO_2$ in MOFs. Attempts to predict maximum loadings in porous materials have been carried out in the past, albeit only for small molecules such as methane, water, or carbon dioxide. Some notable works include the work of Li et al.,[35] on predicting the maximum loading of methane in coal, and Zhao et al.,[36] who used an Artificial Neural Network using structural and energy descriptors to predict maximum loadings of propylene in zeolites. Some works also address the maximum loading by predicting the entire adsorption isotherm, such as that of Chakraborty et al.[37,38] on the full isotherms of methane, $CO_2$, and $N_2$ in zeolites. Despite the substantial amount of literature available on the topic of adsorption in porous materials, little research so far has been performed on using such ML methods for large alkanes, both linear and branched, in zeolites. One reason for this is the strong focus on gas separation and filtration applications involving smaller adsorbates (such as methane, water, and $CO_2$) and the complexity associated with large branched hydrocarbons.

In this work, a modeling framework using ML models to predict the maximum loadings of both linear and branched alkanes in zeolites is presented. The proposed approach uses two models in a series arrangement (otherwise known as a cascading model). This model splits the task of predicting saturation loadings into two smaller tasks: a classification predicting whether adsorption for a given alkane can fit into the zeolite structure as a binary outcome. If adsorption can occur, then a regression is made to estimate the value of the maximum loading. The classification aspect is handled by a gradient-boosting tree classifier (GBC).[39,40] This model is based on a collection of decision trees in series, where the next tree learns to correct the errors of the previous one.[39] A separate stage composed of a regressor assesses how much the maximum loading is if adsorption indeed occurs. Multiple models are considered, including support vector regressors (SVR),[39,41] gradient-boosted tree regressors (GBR),[39,40] extreme gradient-boosted regressor (XGB),[42] and the TabPFN model,[43] which is a prefitted transformer model. Except for TabPFN, all models are also sensitive to hyperparameters, such as the learning rate (strength of error correction) for the GBC or the $\varepsilon$ parameter (prediction tolerance) in the SVR. To ensure the highest performance for any data set, hyperparameters are determined through Bayesian optimization.[44]

To generate training and testing data for maximum loadings, the molecular simulation software RASPA2[45,46] is used. RASPA2 is open-source software designed to compute adsorption and diffusion in porous materials such as zeolites using force-field-based molecular simulation. To compute adsorption isotherms, grand-canonical Monte Carlo ($\mu VT$) simulations are performed using molecular interactions described by the TraPPE[47,48] and TraPPE-zeo[49] force fields. With these and the available Configurational-Bias Monte Carlo



https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314

(CBMC) algorithm[13,48] to generate conformations of large alkanes, RASPA2 provides accurate predictions for adsorption isotherm data.[50,51] In this work, only alkanes with up to 16 carbon atoms (both linear and branched) are considered, since these are the largest alkanes encountered in kerosene.[52] Furthermore, only side groups going up to 3 carbon atoms (i.e., methyl, ethyl, propyl, and isopropyl groups) are considered. An important limitation of the analyzed zeolites is that only full-silica frameworks are considered. This leads to a final population of 45 alkanes and 97 zeolites, respectively, resulting in a total of 4365 data points.

The paper is structured as follows. In Section 2, the methods used to gather the necessary data (for both the individual components and the maximum loading), to create the model, and to evaluate its performance are explained. In Section 3, the model's performance, interpretability, and robustness are discussed, and a comparison with a deep learning approach is presented. In Section 4, concluding remarks are provided, together with a description of possible improvements and future work.

This article also contains Supporting Information. SI1 is a pdf file containing the list of all zeolites and alkanes considered in this work, as well as the optimal model hyperparameter values and raw results of the robustness study. SI2 is an Excel sheet acting as a log of all RASPA2 simulations (input data for the ML models, results, and parameters). SI3 is a folder containing all of the Python codes used for this work.

## 2. METHODOLOGY

### 2.1. Zeolite and Alkane Selection

The computational space for possible simulations is too large for all of the possible alkane−zeolite combinations. As such, some degree of filtering is needed to limit the number of computations and make the simulations more manageable. Alkanes for this work are selected from a database from previous work.[18] This includes a complete list of all isomers from $n$-$C_1$ to $n$-$C_{20}$. Since the primary focus of this paper is SAF, the scope is restricted to alkanes containing up to 16 carbon atoms, corresponding to the largest molecules typically present in kerosene. Furthermore, this work is limited to alkanes with methyl, ethyl, propyl, and isopropyl side groups, as larger groups will not be able to diffuse inside the zeolite channels.[18] A base set of alkanes comprising linear alkanes from methane to nonane, as well as some small isomers such as isobutane, are selected to form a base set of selected molecules. Larger alkanes and isomers are selected using a random sampling method from a list of all hydrocarbons with a given carbon number, with a bias favoring molecules that are more different than the currently present ones. Each molecule was represented as a feature vector $\mathbf{x}$ comprising the main-chain carbon count, the total carbon number, the number of side groups, and the fraction of carbons not included in the main chain. For a candidate molecule $A$ (represented by vector $\mathbf{A}$), its structural similarity to each alkane $B$ (vector $\mathbf{B}$) in the molecular set was quantified using the Tanimoto distance[53] defined as

$$D_T(A, B) = 1 - \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|^2 + \|\mathbf{B}\|^2 - \mathbf{A}\cdot\mathbf{B}} \quad (1)$$

This is used to quantify how close two molecules are. Only when the average distance $\bar{D}_T$ is above 0.3 for all of the already selected alkanes, the molecule is added to the set. Otherwise, a new candidate is sampled from the database of all isomers. This technique is used to bring the total number of selected molecules to 47. The full list of alkanes can be found in the Supporting Information.

The zeolite data is obtained by combining the IZA online database,[54] the iRASPA visualization software,[55] and the structures found in the Atlas of Zeolite Structures.[56] From these three databases, zeolite structures are selected on the basis of three criteria:

*Orthorhombic unit cells*: To limit the computational load and improve simulation speed, only orthorhombic unit cells are considered. This is because using non-orthorhombic cells requires intermediate steps (such as matrix inversions for computing distances[45,46]) and is thus more computationally expensive.

*Limited or no enclosed voids*: In GCMC simulations, a molecule is inserted where the volume allows it. This means that a molecule can be inserted in a void that is disconnected from the main channel system, which is not physically possible in adsorption experiments. Methods to block these pockets do exist, but it has been chosen to exclude the most extreme cases, such as CFI-type or IRN-type zeolites.

*Nonzero accessible volume* ($AV$): A subset of zeolite structures lacks continuous channels or interconnected pore networks, resulting in negligible accessible volume. Since such frameworks do not support meaningful transport or adsorption behavior, these are excluded from further consideration.

Out of the 232 structures common to both databases, 97 satisfy the above requirements. The complete list of used alkanes and zeolites can be found in the Supporting Information.

### 2.2. Computing the Maximum Loading

The maximum loading of each alkane−zeolite combination is determined by a two-step process. The first step addresses whether adsorption of the alkane is possible in a given framework. For this, the **ZEO++** tool is used.[57] Given a certain probe radius for a given molecule, **ZEO++** can determine whether channels in the framework are accessible or not. If no channels are accessible, it is because the minimum cross-sectional size of the molecule (determined from the minimum ellipsoid diameter obtained in RDKit[58]) cannot be accommodated by the restricting pore diameter (RPD) of the zeolite channel. As such, the maximum loading $q_{max}$ is set to zero. If channels are accessible, **ZEO++** generates a block-pockets file to wall off possible inaccessible pockets in the structure. The maximum loading is then computed in RASPA2 by using these block pockets to prevent molecules from being generated in inaccessible pockets. All Monte Carlo simulations are carried out in the grand-canonical ($\mu VT$) ensemble at a fixed temperature of 298 [K]. To ensure that maximum loadings are encountered, all simulations are carried out at an external pressure of $10^{10}$ [Pa]. An exception is made for simulations of methane ($10^{14}$ [Pa]) due to its smaller size. The high fugacities are used only to approach saturation and do not represent experimental conditions. In ref 59, it is explained that the maximum loading for systems without hard-sphere interactions should be defined operationally. The fugacity coefficient of the simulations is set to one so that the imposed pressure is equal to the fugacity. The TraPPE[47,48] and TraPPE-zeo[49] force fields are used to describe the intramolecular bonded (bond,



https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314

bending, and torsion) and nonbonded (Lennard-Jones) interactions, as well as guest−guest and guest−host interactions (Lennard-Jones). The zeolite structures are treated as rigid since structural changes during adsorption processes have a minor effect on adsorption.[60] A cutoff radius of 12.8 [Å] is used for describing Lennard-Jones potentials, with tail corrections applied.[13] The alkanes are represented using the united-atom approach (i.e., carbon and hydrogens are represented as a single interaction site) and are inserted into the system using the Configurational-Bias Monte Carlo algorithm.[13,61] The ideal gas Rosenbluth weights of all the considered molecules are computed in separate *NVT* simulations. At high loadings, the acceptance probability of insertion and deletion trial moves becomes very low. In such cases, ensemble averages are no longer representative of the maximum loading. Therefore, the mode of the maximum loadings is used to have a representative result while still retaining statistical robustness. To this end, simulations were run for $2 \times 10^6$ Monte Carlo cycles. The number of trial moves per cycle equals the number of molecules, with a minimum of 20. The number of trial positions in the CBMC algorithm is set to 10.

A possible source of error is the size of the simulation box. Molecular simulations can suffer from finite-size effects, in which systems that are too small do not accurately depict the thermodynamic limit. This is best explained by Figure 1.

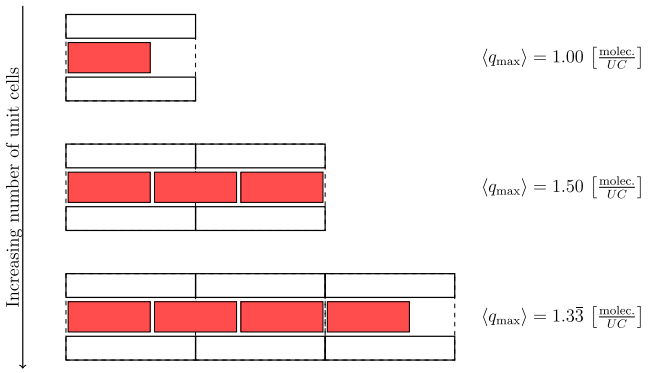



**Figure 1.** Visual depiction of finite-size effects for the adsorption of chain alkanes (red). As both the available space in the zeolite and the size of the molecule are fixed, the maximum loading of possible molecules physically fitting in the channel is capped to an integer amount. This can result in situations where leftover volume may be present, thus influencing the final average of the maximum as a function of system size.

In a simulation box with a fixed number of unit cells through which channels pass (depicted in black), the maximum loading corresponds to the largest integer number of molecules (approximately red rectangles) that can physically fit in the channel. Since the domain is finite, there is a discrete number of possible "slots" in which the molecule can be placed. This has the consequence of artificially induced stepwise behavior in the average loading as a function of system size, which can be mitigated by adding more unit cells in the relevant direction. For this reason, multiple simulation box sizes are used, starting from the smallest possible cell until either a cap on the number of unit cells or a maximum number of unit-cell expansions is reached. This implies that the size of each simulation box is proper for each alkane−zeolite combination. A complete simulation log dictating simulation parameters (e.g., size of the simulation box, simulation time, convergence of finite-size effects, etc.) and values of the resulting maximum loading is available in the Supporting Information.

### 2.3. Descriptor Engineering and Preprocessing

To characterize different alkanes and zeolites, a set of quantities is chosen to describe individual properties, as well as interactions with each other. Since the maximum loading is governed by physisorption, most descriptors used are geometrical or topological in nature. In this study, zeolites and alkanes are both characterized independently using 13 descriptors, with an additional 3 descriptors to introduce physics in the data set. It is intuitive that the maximum loading is dependent on the volume, surface area, and topological properties of a given framework. Topological descriptors include framework density (FD) and topological density on 10-membered rings ($TD_{10}$). To compute the volume and surface properties of the zeolites, the framework is imported into **ZEO++**.[57] Methane was selected as the probe instead of helium to provide a more realistic estimate of the volumes and surfaces available to the molecules, since the smaller diameter of helium will incorrectly count hard-to-reach or inaccessible pores. The void fraction was computed using RASPA2, where methane molecules are inserted according to the Widom particle insertion method.[13] Basic alkane descriptors are computed using the Python RDKit toolbox.[58] This includes molecular descriptors, such as molar mass, number of side groups of each type, total and main-chain carbon counts, and molecular volume. As it is known that zeolites exhibit shape-based selectivity,[9,62,63] an emphasis on shape-based descriptors is made. This includes the acentric factor $\omega$, which is computed by the method developed in the Supporting Information of ref 64. Another essential descriptor codifying the shape is the principal mass moments of inertia. Since RASPA2 generates a Boltzmann distribution of conformers of a molecule, the inertia tensor $\mathbf{I}$ can be built. From these, the values of the principal mass moments of inertia can be obtained by computing the eigenvalues of the tensor.

As mentioned earlier, the main reason molecules are not adsorbed in the zeolite is because of steric hindrance. As such, an essential dimension is an effective diameter for the molecule. Consequently, a proxy is derived by encapsulating the molecule within an ellipsoid to obtain an approximate of the size. Since larger molecules tend to be more flexible, this process is repeated for a large number of conformers created in the ideal gas phase with RASPA2. The energies of each of these conformers are computed using the UFF[65] energy calculator in RDKit so that results can be Boltzmann-weighted. The UFF calculator computes the potential energy of an isolated molecule. This method is used to obtain the minimum and maximum diameters of the resulting ellipsoid, for which the mol-ellipsoid package[66] is used.

To enhance the information given by the descriptors, meaningful ratios can be formulated to emphasize interactions between the two. One of these values is the roughness factor RI of zeolite channels, defined as

$$\text{RI} = \frac{\text{AS}}{\text{AV}^{2/3}} \tag{2}$$

in which AS and AV are the methane-probed available surface and available volume inside the zeolite, respectively. The motivation behind this is that dimples in the channels can result in preferential adsorption sites for some molecules.


https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314

Composite descriptors are also introduced to capture physical and mutual interaction relationships, thus creating a hybrid machine learning model that is physics-informed on top of being data-driven.[38] An example of this is the ratio $\chi$ between the minimum molecular diameter and the restricting pore diameter

$$\chi = \frac{D_{\text{Min}}^{\text{Mol}}}{\text{RPD}} \tag{3}$$

This helps encode the main steric relationship directly in the data, which would help the model understand it directly. It is using the same reasoning that a packing factor between zeolite and molecular volumes

$$\Psi = \frac{\text{AV}}{V_{\text{Mol}}} \tag{4}$$

is introduced. As maximum loading is physisorption-dependent, multiple layers of adsorbate can form, which means that the volume relationship between adsorbate and adsorbent becomes important since it constitutes a physical constraint. A summary of all descriptors used for the models is shown in Table 1.

Both the maximum loading and the descriptors naturally span multiple orders of magnitude. This can result in skewed distributions, which may result in weak performance from the models due to the high variance.[39,67] This step is especially necessary since SVM data needs to be scaled due to its reliance on distances to target data points.[41] As such, a check of the data spread needs to be carried out. This can be addressed by making a probability distribution of the normalized values of the maximum loading, as shown in Figure 2.

It is clearly observed that under nontransformed conditions, the data presents a right-tail skew, i.e., most values cluster on the lower end, with only a few reaching the high end. Since unusually large values of maximum loadings can be observed for special cases of methane in large-pore zeolites, and larger molecules generally show lower loadings, a logarithmic transformation is applied to compress the dynamic range of the data. This operation reduces the relative influence of high-capacity outliers while expanding the spread to smaller values, thereby improving the sensitivity of the model for the full range of maximum loadings.[68] As such, the final transformation carried out on the maximum loading is given by

$$q_{\text{max}}^{*} = \log_{10}\left(1 + \frac{q_{\text{max}}}{q_0}\right) \tag{5}$$

with $q_0$ being a unit maximum loading at 1 [molec./m$^3$] to make the argument inside the logarithm dimensionless. This results in a distribution that more closely resembles a bimodal distribution, thus making it less prone to biases. This helps preserve the physical meaningfulness of the zero values of adsorption in the final data set. The same transformation is carried out for highly skewed descriptors, such as the accessible surface area and volume of the zeolite, as well as the principal mass moments of inertia of the molecules.

## 2.4. Machine Learning Models

In the majority of the available literature, only a single ML model is used to predict the selected target. However, early attempts at using only a single model have resulted in overall commendable performance, but with many data points that are incorrectly predicted to be cases of positive adsorption, i.e., $q_{\text{max}} > 0$. To this end, the proposed model is a cascading ML model, i.e., a model made up of two ML models placed in series. The motivation behind this choice is physical. The first model tries to predict whether adsorption is likely to occur as a binary outcome. If it is, then the regressor will predict the actual maximum loading. Otherwise, the maximum loading $q_{\text{max}}$ can be set to 0. This means that the classifier is trained on both data points of zero adsorption and nonzero adsorption, while the regressor is trained on nonzero adsorption data points only. Different models are considered for the second step. These include (extreme) gradient-boosted trees, support vector machines, and the TabPFN model. As all of these models are already discussed in detail in the literature, only a short summary is provided here.

Gradient-Boosted Trees (GBT)[39,40] are an ensemble of decision trees that are trained sequentially. Since a singular decision tree runs the risk of overfitting data, multiple trees are

**Table 1. Table Containing All Zeolites, Molecules, and Composite Descriptors Used in the Model[a]**

| Descriptor | Category | Symbol | Units |
|---|---|---|---|
| Void fraction | Zeolite | $\phi$ | [-] |
| Methane-accessible area | Zeolite | ASA | m$^2$/g |
| Methane-accessible volume | Zeolite | AV | cm$^3$/g |
| Overall accessible volume | Zeolite | AccV | % |
| Methane-nonaccessible area | Zeolite | NASA | m$^2$/g |
| Methane-nonaccessible volume | Zeolite | NAV | cm$^3$/g |
| Framework density | Zeolite | FD | g/cm$^3$ |
| Topological density (10-membered rings) | Zeolite | $\text{TD}_{10}$ | – |
| Channel dimension | Zeolite | $d$ | Å |
| Restricting pore diameter | Zeolite | RPD | Å |
| Largest cavity diameter | Zeolite | LCD | Å |
| Molar mass | Zeolite | MM | g/mol |
| Gravimetric density | Zeolite | $\rho$ | g/cm$^3$ |
| Molecular weight | Alkane | MW | g/mol |
| Main chain length | Alkane | MCL | # of carbons |
| Total carbon count | Alkane | TCC | # of carbons |
| Side group composition | Alkane | $[n_{\text{C}}, n_{\text{C2}}, n_{\text{C3}}, n_{\text{iC3}}]$ | counts |
| Acentric factor | Alkane | $\omega$ | – |
| Molecular volume | Alkane | $V_{\text{M}}$ | Å$^3$ |
| Minimum ellipsoid diameter | Alkane | $D_{\text{Min}}^{\text{Mol}}$ | Å |
| Maximum ellipsoid diameter | Alkane | $D_{\text{Max}}^{\text{Mol}}$ | Å |
| Principal mass moments of inertia | Alkane | $\text{PMI}_i$, $i = \{1, 2, 3\}$ | amu·Å$^2$ |
| Ratios of first and second principal mass moments of inertia | Alkane | $\frac{\text{PMI1}}{\text{PMI2}}$ | [-] |
| Ratios of second and third principal mass moments of inertia | Alkane | $\frac{\text{PMI2}}{\text{PMI3}}$ | [-] |
| Roughness index | Composite | RI (see eq 2) | [-] |
| $\frac{D_{\text{Min}}^{\text{Mol}}}{\text{RPD}}$ | Composite | $\chi$ (see eq 3) | [-] |
| $\frac{\text{AV}}{V_{\text{Mol}}}$ | Composite | $\Psi$ (see eq 4) | [-] |

[a]Values of descriptors for the zeolites were obtained from refs [54,55,56]. Values of descriptors for the molecules were obtained using the RDKit library in Python.[58] The definitions of RI, $\chi$, and $\Psi$ can be found in eqs 2, 3, and 4, respectively.



https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314

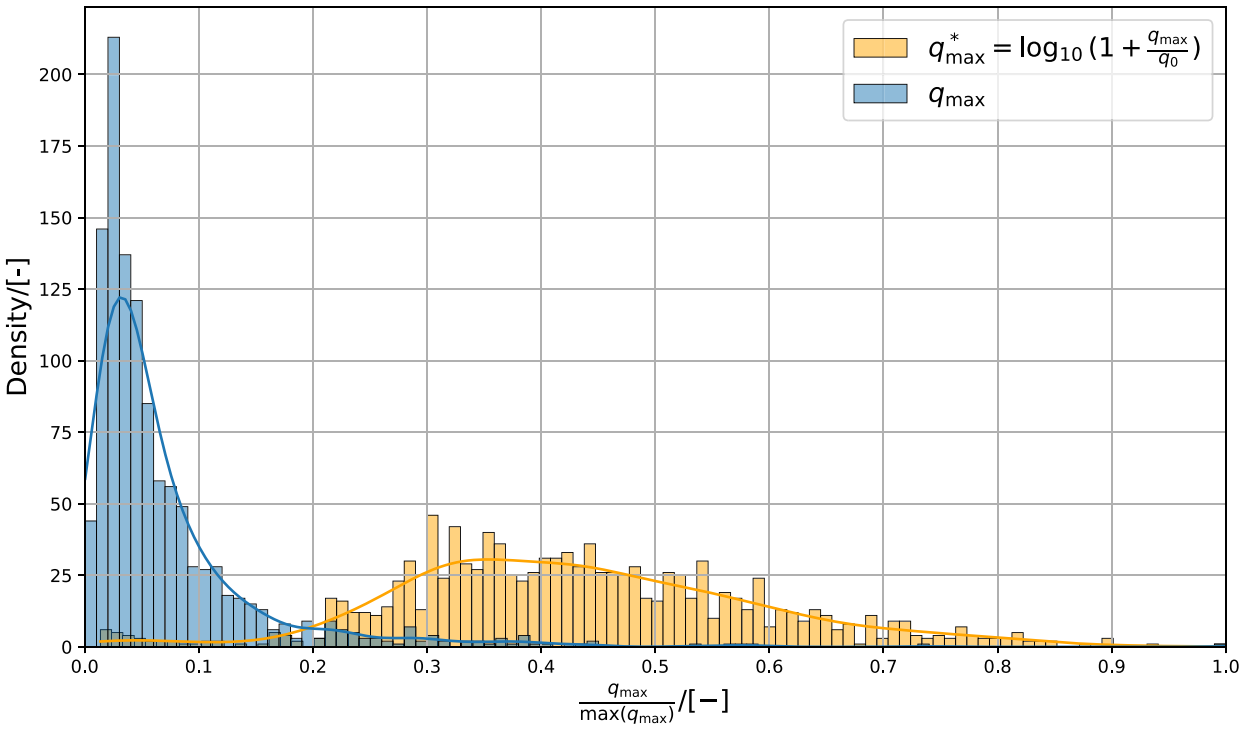



**Figure 2.** Probability distribution of the (nonzero) values of the maximum loading (nontransformed $q_{max}$ in blue, transformed $q_{max}$ in yellow). The blue distribution presents a right tail, which poses the risk of biasing. To deal with this, the yellow distribution is generated, resulting in much better distributed data.

created. The algorithm works by making an ensemble of decision trees, where an initial tree is trained, and each tree made after this one tries to correct the residuals of the previous one. When constructing the next tree in the ensemble, pseudoresiduals, defined as the negative derivative of a specified loss function evaluated on the current one, are determined. A tree regressor is fitted on top of these pseudoresiduals and is then added to the ensemble. The final model prediction is thus a combination of the initial weak performer added with contributions from each of the trees trained on pseudoresiduals. A variant of this model is the extreme gradient-boosted tree (XGB).[42] Traditional gradient boosting is only considered at the first derivative of the loss function. To strengthen the boosting, XGB determines the next tree in the ensemble by minimizing an objective using not only the first and second derivatives of the original loss function but also a term penalizing the pseudoresidual tree complexity (via L1 and L2 regularization). Support Vector Regressors (SVR)[39,41] work by optimizing a function across a multidimensional space that does not deviate more than a specified $\varepsilon$ margin, while minimizing model complexity. To allow these models to not compute the mapping directly, support vector regressors use kernel functions to represent the nonlinear relationships. Unlike the GBT and XGB methods, SVR is sensitive to data scaling due to its distance-based method. TabPFN[43,69] is a Prior-Data Fitted Network trained over millions of synthetic tabular tasks, over which it learns an approximate to Bayesian inference. For regression tasks, it outputs a point estimate of the value from a predicted probability distribution of the target. This means that the training data are merely used to allow the model to condition the right prior. Because it is a pretrained model, it does not need any hyperparameter tuning. Some of these models are subjected to a set of hyperparameters that determine their training dynamics, whose combinations can result in better or worse performance. To find an optimal set of hyperparameters, Bayesian optimization[70] is used. It aims to maximize a given objective by means of a probabilistic surrogate and an acquisition function, guiding the optimizer. During each iteration, a value of the objective function is sampled at a certain point, which helps in actively updating a map of the optimization space by being fed to the surrogate. In parallel, an acquisition function proposes a point that would result in the largest expected improvement, where the objective is resampled. This method is a common approach to optimizing model hyperparameters,[71] as it offers a more informed optimization than random or grid search, with an accuracy comparable to more complex methods, while still being computationally light. For this work, both the optimization of the classifier and regressor were given 150 iterations, with 10% dedicated to build the initial surrogate model. In this work, all models are implemented by using Python. Both the GBT and SVM are implemented using scikit-learn,[72] XGB is implemented using XGBoost,[42] and TabPFN is implemented through the TabPFN Python package.[43,69]

## 2.5. Analysis Metrics

To quantitatively assess model performance, several performance metrics need to be established.[39] Since the cascade is a combination of a classifier and a regressor, different sets of metrics need to be defined. Classifier performance is usually assessed using accuracy. For a set of outcomes made up of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), the accuracy is defined as

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{6}$$



https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314

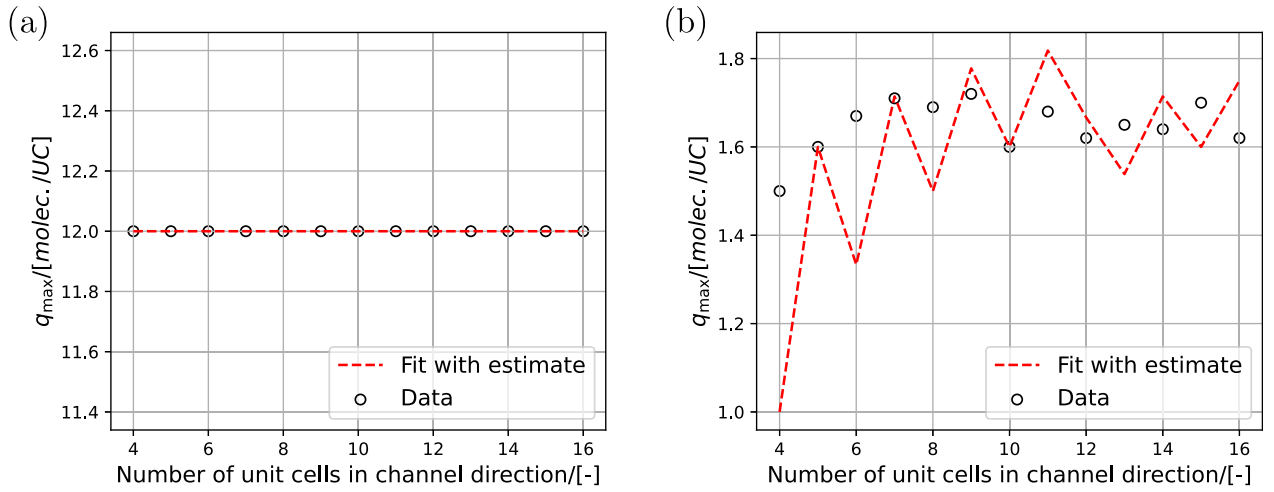



**Figure 3.** (a) Finite-size effects for the adsorption of ethane in AET-type zeolite. As ethane is much smaller than the channels of AET-type zeolite, the maximum loading changes only slightly as the simulation box increases. (b) Finite-size effects for adsorption of 2-methyl-4-ethyl-nonane in SAF-type zeolite. Larger molecules in narrower channels show stronger variations before converging (here around ∼1.65 molec./UC). The red line shows a fit to eq 13.

This metric could be misleading if there is a class imbalance in the data, which may not reflect poor prediction performance on the minority class. For this purpose, the class-wise $f_1$-score is introduced. It represents how reliably a model can identify a class without missing real cases or raising false alarms. It is defined as

$$f_1 = 2 \times \frac{P \times R}{P + R} \tag{7}$$

with precision $P$ and recall $R$, both given as

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{8}$$

It allows both minimization of the amount of false positives (via precision) and maximization of the capture of true positives (via recall). By averaging the score between classes, a MACRO-$f_1$

$$\text{MACRO-}f_1 = \frac{1}{N}\sum_{c=1}^{N} f_{1_c} \tag{9}$$

is created, ensuring that the final metric of the dominant and minority classes are fairly represented. In eq 9, $N$ is the total number of classes. This advantage makes the MACRO-$f_1$ score a suitable choice as an objective function for Bayesian optimization of the classifier. Performance of the regressor is primarily assessed using the coefficient of determination ($R^2$), the mean absolute error (MAE), and the mean squared error (MSE). For a set of true values $y_i$ with mean $\overline{y}_i$ and a set of predictions $\hat{y}_i$, both having $N$ samples, these metrics are defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \overline{y}_i)^2} \tag{10}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}| \tag{11}$$

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{12}$$

The mean square error is used as an objective function for optimizing the regressor, as differences are punished more severely while keeping the objective differentiable. To ensure a robust and unbiased estimate, the selected objectives are the result of a 5-fold cross-validation on the training set. This procedure splits the training set into 5 subsets ("folds"). The model is trained on 4 folds and tested on the fifth one. This is repeated so that each fold is used as a test once, and the average across all folds is reported.

## 3. RESULTS

### 3.1. Analysis of Finite-Size Effects

A closer look is taken at the simulation results with RASPA2 to investigate to what extent finite-size effects are present. Two typical examples are presented in Figure 3.

This figure shows the origin of finite-size effects and how the final result of the maximum loading is influenced. Figure 3a shows the finite-size effects of ethane in the AET-type zeolite. Since ethane is small compared to the channel size, there is no variability in the final result as more unit cells are added. In sharp contrast, Figure 3b, which shows 2-methyl-4-ethyl-nonane in SAF-type zeolite, demonstrates that the smallest simulation box (4 unit cells) does not produce a representative result. Since 2-methyl-4-ethyl-nonane is a relatively large alkane and the SAF-type zeolite has short channel lengths (8.3173 Å unit cell dimension along the channel[56]), a larger number of unit cells is required to obtain a reliable ensemble average, as demonstrated by the convergence of the maximum loading with an increasing number of unit cells.

If one would use the principle behind Figure 1, then one could be compelled to fit a function to see how many unit cells in the channel direction are needed to obtain the result of an infinitely large domain. This can be approximated using the unit cell channel length $L$, the number of unit cells in the channel direction $x$, and a fitted parameter $a$ resembling assumed constant size of the molecule. The resulting function would be



https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314

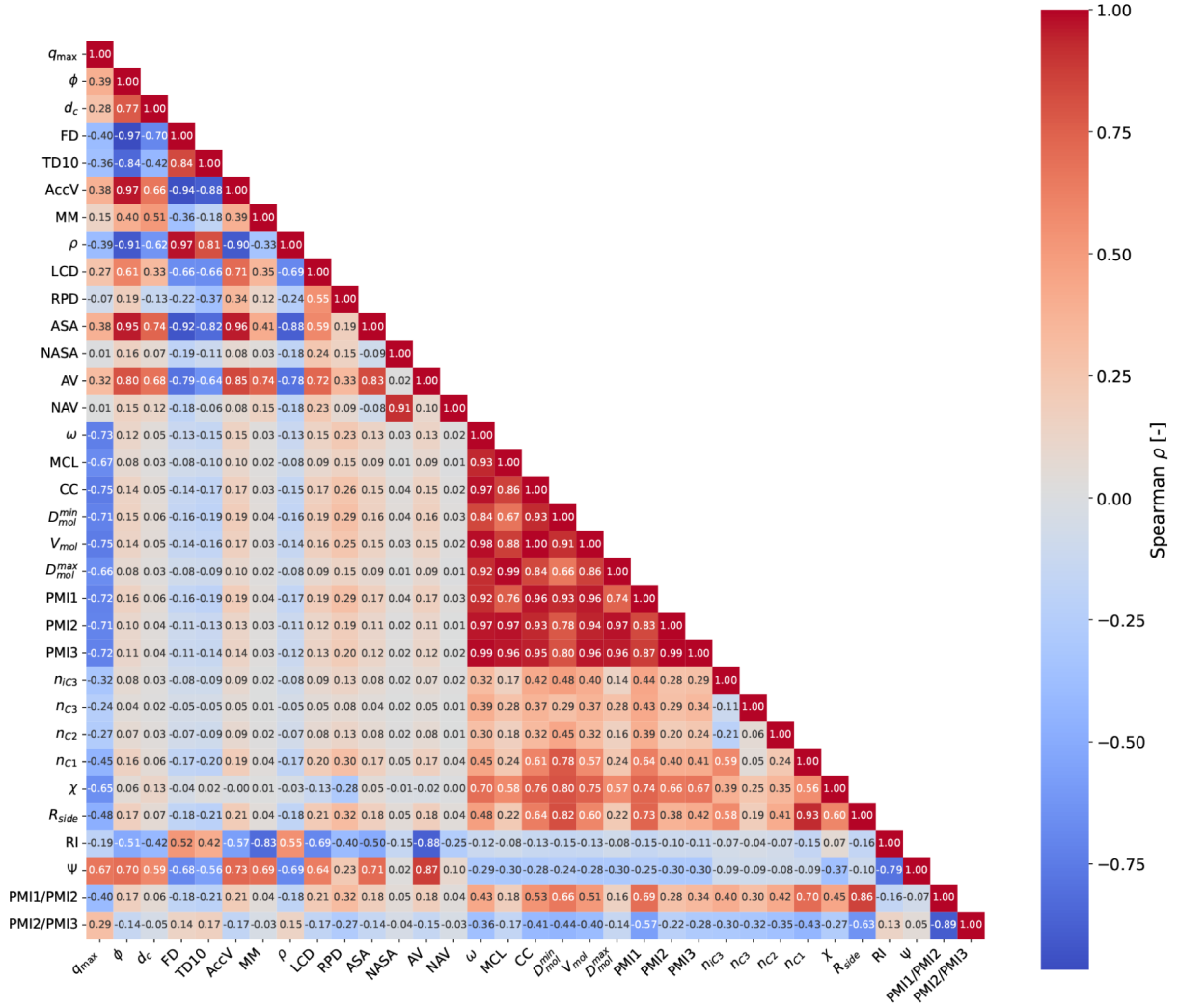



**Figure 4.** Spearman Correlation Matrix between descriptors and $q_{max}$. Despite the fact that the molecule and zeolite descriptors exhibit high internal correlation with each other, it is clear that both are essential to predicting the maximum loading. Molecular descriptors appear to provide the dominant contribution, while zeolite-based descriptors add complementary information. All descriptors in this matrix are defined in Table 1.

$$E(a, L, x) = \frac{\text{int}(xL/a)}{x} \tag{13}$$

where the function int($x$) rounds its argument to the lowest integer in the limit of $x \to \infty$, $E(a,L) = L/a$, as expected. The fit (eq 13) is applied to the data of Figure 3a and 3b, as shown by the dashed red line. In Figure 3b, it can be seen that the general trend is reproduced, but the nonlinearity introduced by the numerator of eq 13 leads to visible instability in the result. This illustrates the complexity of the problem at hand as well as proves that the simple model of eq 13 is insufficient to predict the maximum loading of hydrocarbons in zeolites. This supports the use of machine learning for this specific task.

### 3.2. Correlation Analysis of Descriptors

To confirm the choice of descriptors, a correlation matrix can be constructed. Using a correlation coefficient, one can assess whether a feature influences the maximum loading or whether it is strongly correlated with other features, thus making it redundant. Because adsorption inherently exhibits nonlinear behavior, the Spearman $\rho_S$ coefficient[73] is used, as it can capture monotonic nonlinear relationships and is more robust to outliers than the Pearson correlation coefficient. A correlation between the descriptors themselves, as well as with the maximum loading, is presented in Figure 4. Only points with nonzero $q_{max}$ are considered for these correlations.

Selecting descriptors that are correlated to the target helps the model capture clear, relevant patterns, possibly enhancing the performance. However, correlated descriptors introduce redundancy, as such features tend to provide overlapping information. The latter may negatively impact the model because of multicollinearity while also making the model more complex by addition of a new dimension. The first cluster that can be noticed is the one of zeolite descriptors in the top-left ($\phi$, ASA, AVA, AccV, FD, $TD_{10}$, $d$, RPD, LCD, $\rho$). From there, it can be noted that volume- and topology-based descriptors are moderately correlated to $q_{max}$. This means that despite not solely determining $q_{max}$, these descriptors still impose some level of constraint. The void fraction, accessible area, accessible volumes, framework density, topological density, and gravimetric density are highly correlated ($\rho_S > 0.9$). This is expected, as the descriptors are mutually dependent on each other. These descriptors also moderately correlate to the maximum loading, as larger channels increase volumes and areas while driving densities down. It is important to note the weak correlation between RPD and $q_{max}$. This is expected because the steric exclusion effects are primarily handled in the



https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314

classification stage of the cascade model, which determines whether adsorption can occur. The regression analysis therefore focuses only on cases in which molecules already fit inside the pore system.

Molecule descriptors exhibit stronger correlations to $q_{max}$, as well as between each other, due to all descriptors being directly related to the size and shape of the molecule. Shape-driven descriptors (such as the principal moments of inertia and the acentric factor) are highly correlated to the maximum loading, hinting that shape-based selectivity may be inherently present in the data set. Correlations involving the branching of the alkanes are seen to be low to moderate, and as such are expected to have limited impact. This is mostly because descriptors pertaining to the size of the molecules already include the information. Overall, it can be concluded from the Spearman correlation matrix in Figure 4 that the extent of the maximum loading is dependent on both zeolite and molecule descriptors, although there seems to already be a dominance of the latter in the final result. This is expected as these are known samples that get adsorbed, so there is no "exclusion regime" that needs to be defined with pore geometry, and as such will operate more as a soft upper bound. This allows molecular variation, which is expected to dominate here. It is also observed that a lot of descriptors are correlated with each other. This in itself is not fatal but indicates that some of these descriptors may be removed if found to be of little importance to the model prediction.

### 3.3. Model Performance and Outlier Analysis

Different metrics are relevant to assess the performance of each stage of the cascade model. As such, each element of the cascade is assessed individually. The classifier's performance is quantified by means of a confusion matrix and a performance table. Regressor's performance is addressed by means of parity plots. All results are obtained using a fixed random seed and with 20% of the data withheld for testing to obtain consistent results.

The classifier confusion matrix and performance table obtained from the classifier can be found in Figure 5 and Table 2, respectively. As can be seen in the confusion matrix, the classifier has excellent performance with near-perfect accuracy. This can be attributed to Bayesian optimization finding an optimal set of training hyperparameters. From Table 2, it can be observed that on a per-class basis, both have very high precision and recall. Despite this, the former is higher for adsorption cases, and the latter is higher for nonadsorption cases. This suggests that the classifier is more conservative, missing true adsorption cases (lower recall) rather than predicting adsorption happens when it should not happen (lower precision). This is further supported by the results in the confusion matrix. It can also be observed that the $f_1$ scores of both classes are very close to each other, hinting that the model can predict both adsorption and nonadsorption cases with comparable strength. As shown in Figure 5, only five samples were incorrectly labeled. These are given in Table 3, with their true and predicted labels.

Table 3 shows that there is not one molecule or zeolite particularly prone to being mislabeled. Closer analysis of these outliers reveals that UWY/4-isopropyl-3,4,4,5-tetramethyloctane is missed by the classifier because it is the adsorbing combination with the highest value of the minimum molecular diameter-to-restricting pore diameter ratio $\chi$ (~1.55). Since the combination with the second-highest value, BEC/4-isopropyl-3,4,4,5-tetramethyloctane is correctly labeled, suggesting that there is a threshold value of $\chi$ in the model beyond which adsorption is predicted as not happening. This is grounded in physics, as the diameter-to-restricting pore diameter ratio accounts for the steric hindrance of molecules in the zeolite channels. A similar explanation applies to VET/3-ethyl-2,2,6-trimethyloctane and AFO/*n*-nonane, which both have the lowest values of the packing factor Ψ for adsorbing

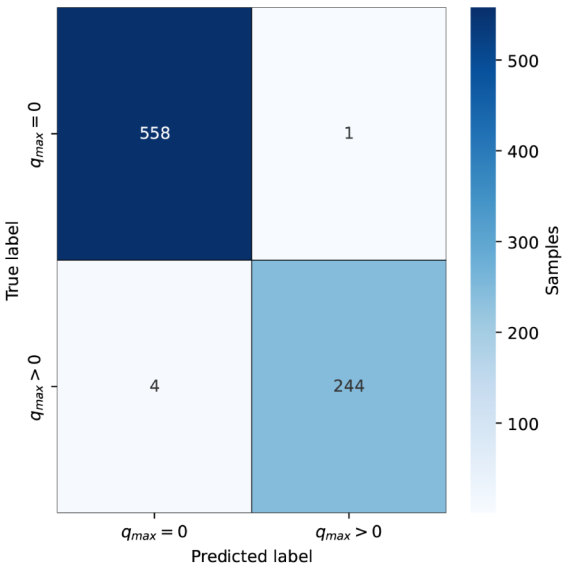



**Figure 5.** Gradient-boosted classifier confusion matrix. Each cell represents the number of samples. The low number of total mislabeled points reflects near-perfect accuracy. The model is slightly more prone to predicting zero adsorption for combinations that would adsorb rather than incorrectly predicting adsorption where none should occur. These values are used to compute the classification performance metrics, namely accuracy, the class-wise $f_1$-score, and the MACRO-$f_1$ score using eqs 6, 7, and 9, respectively.

**Table 2. Summary Table of Classifier Performance Metrics (eqs 6, 7, 8, 9) for Predicting Adsorption and Nonadsorption Cases, Both Overall and Per-Class Basis[a]**

| Class | Metric | Value |
|---|---|---|
| Overall | Accuracy | 0.9938 |
| $q_{max} = 0$ | Precision | 0.9929 |
| | Recall | 0.9982 |
| | $f_1$-score | 0.9955 |
| $q_{max} > 0$ | Precision | 0.9959 |
| | Recall | 0.9840 |
| | $f_1$-score | 0.9899 |

[a]The high accuracy demonstrates that the model reliably predicts the correct adsorption case. Class-wise precision, recall, and $f_1$ scores show that the model is more likely to predict no adsorption when adsorption should occur rather than vice versa.

**Table 3. Incorrectly Predicted Alkane–Zeolite Combinations (Test Set)**

| Zeolite | Molecule | Predicted label | True label |
|---|---|---|---|
| NES | *n*-Butane | $q_{max} > 0$ | $q_{max} = 0$ |
| AFO | *n*-Nonane | $q_{max} = 0$ | $q_{max} > 0$ |
| MRE | 2-Methylpentane | $q_{max} = 0$ | $q_{max} > 0$ |
| UWY | 4-Isopropyl-3,4,4,5-tetramethyloctane | $q_{max} = 0$ | $q_{max} > 0$ |
| VET | 3-Ethyl-2,2,6-trimethyloctane | $q_{max} = 0$ | $q_{max} > 0$ |



https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314

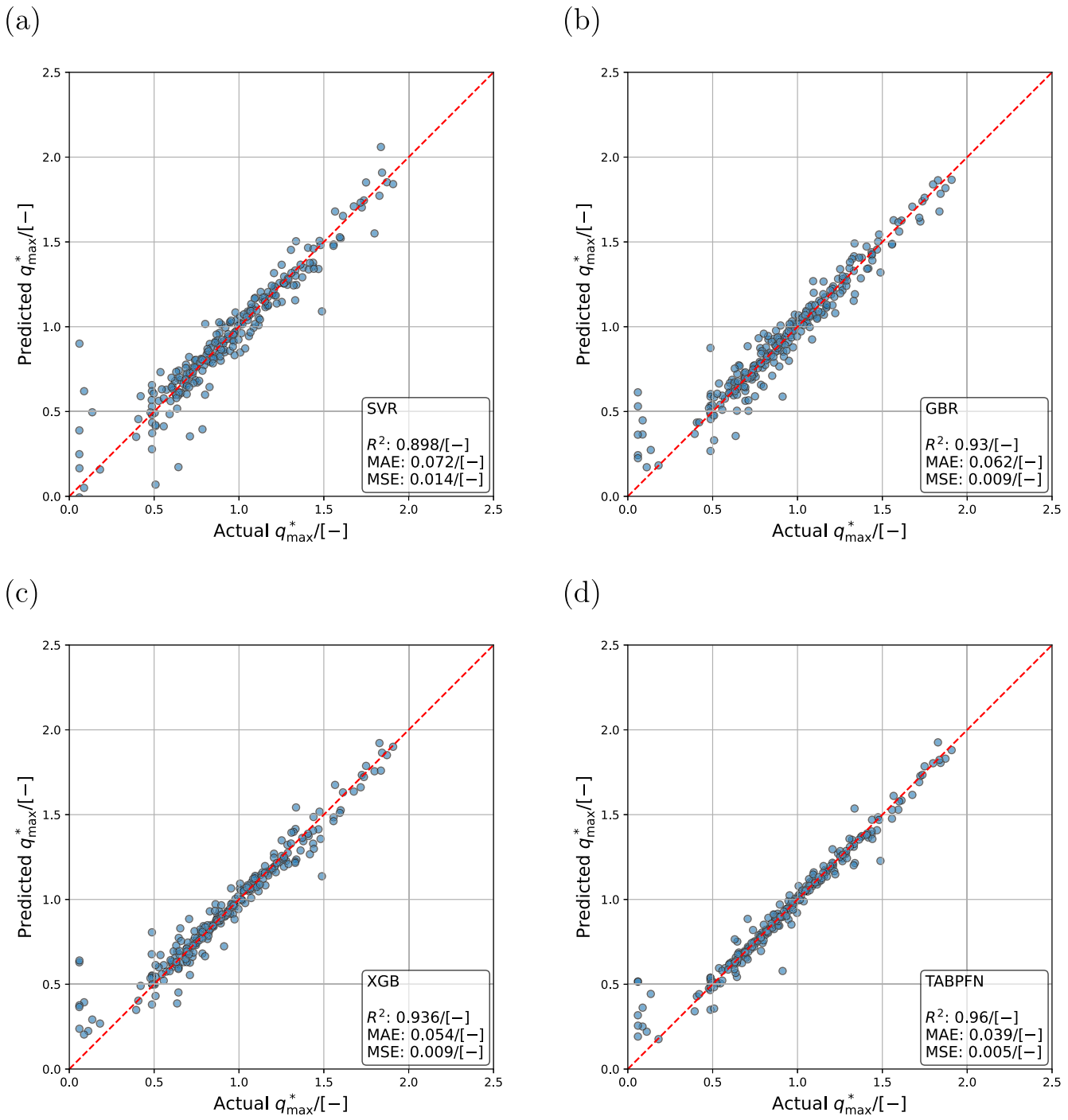



**Figure 6.** Parity plots for all four considered regressor models. (a) Support Vector Regressor, (b) Gradient Boosting Regressor, (c) Extreme Gradient Boosting, and (d) Tabular Prior-data Fitted Network with their associated performance metrics. All models exhibit high performance, though the very low maximum loading regime dominated by larger and more bulky molecules remains somewhat challenging. SVM lags behind mostly due to the high dimensionality of the problem. GBR and XGB perform well, while TabPFN achieves the best performance by capturing dependencies that the other models cannot.

cases, thus also hinting at the presence of a volume-based threshold. As such, a possible strategy to improve the model would be to generate more training data close to these islands of data. For the regressor stage, each of the models is trained and fitted to the maximum loadings obtained from RASPA2. The parity plots for each of the regressors are presented in Figure 6. This includes the respective values of $R^2$, MAE, and MSE. In all plots, the ideal fit line ($q_{max}$(predicted) = $q_{max}$(actual)) is depicted by using a red dashed line. A slight overprediction at low maximum loadings is observed for all models, visible in the lower-left region of the parity plots (Figure 6). This region corresponds to highly confined zeolites combined with large, branched alkanes, where adsorption is determined by strong steric constraints. Under such conditions, small variations in channel dimensions or molecular shape can produce disproportionately large changes in the number of accessible configurations, leading to a sharp transition between fitting and not fitting of adsorbate



https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314

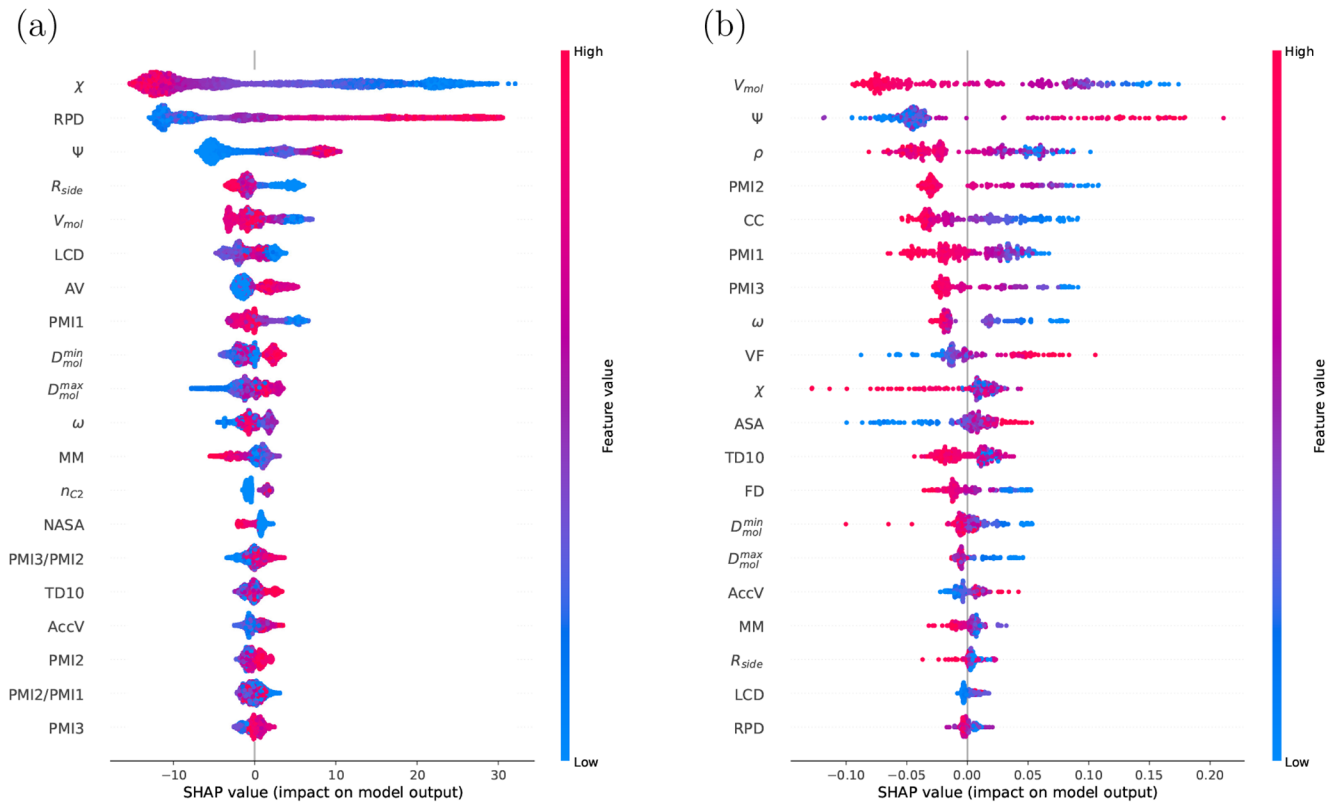



**Figure 7.** (a) SHAP value distribution of the classifier. $\chi$ and RPD dominate the outcome, showing that the model captures the steric hindrance in the zeolite channels. Important molecule-based descriptors include PMI1 and $R_{side}$, indicating that the model accounts for diffusion limitations of bulkier molecules. (b) SHAP value distribution of the regressor. Volume-based constraints dominate predictions via $V_{mol}$, Ψ, and $\rho$, with contributions from molecular shape descriptors such as principal moments of inertia, $\omega$, and $D_{mol}^{min}/D_{mol}^{max}$, as well as $R_{side}$.

molecules. Predictions become particularly sensitive to uncertainties in geometric descriptors, such as ellipsoidal diameters or restricting pore diameters. This behavior resembles the activity-cliff effects reported in our previous work on Henry coefficients.[18] These confined systems represent a small fraction of the data set and span a narrow range of $q_{max}$, which increases mean absolute and mean square errors, especially after the logarithmic transformation.

Out of all models, TabPFN performs the best, with the highest $R^2$ and lowest errors (both MAE and MSE). This is followed by XGB, GBT, and SVM. The SVM achieves decent accuracy but is known to struggle with complex nonlinear dependencies and high-dimensional feature interactions, leading to systematic underprediction for larger or geometrically intricate molecules. Since the tree-based models are overperforming by some margin, it could also indicate that the decision boundary the model needed to fit had overlap zones, which is a known issue for SVRs.[39,41] The tree-based methods (GBR and XGB), due to their boosting capabilities, can use the residuals to create better trees. This explains the increased performance, with XGB being slightly better due to the built-in mechanisms that the GBR does not have (e.g., regularization and depth-first pruning). Lastly, TabPFN exhibits the best metrics due to its transformer-based architecture, allowing it to capture more complex relationships that would otherwise not be captured.

### 3.4. Model Interpretability

ML models are considered black boxes, as it is hard to understand the decisions made by a model solely by looking at its parameters.[74] This can be detrimental to the reliability and scientific relevance of the model, as it may not confirm whether the model has learned the intended pattern. SHapley Additive exPlanation (SHAP) analysis can help confirm this.[75] Based on the Shapley values from game theory,[74] this method offers a way to quantify how much each feature contributes to the final prediction of the model. It is important to note that the SHAP values themselves carry no meaning, as only the sign of the resulting values (whether it pushes the prediction up or down) is of interest. The SHAP value distributions of both classifier and regressor models are presented in Figure 7. These results are both obtained by fixing the random seed and using the GBT regressor, as it is the highest accuracy model that could work with SHAP.

It is clear from Figure 7a that the classifier relies mostly on steric constraints when predicting the likelihood of adsorption. This is shown by the ratio between the minimum molecular diameter and zeolite, with the restricting pore diameter $\chi$ being the most influential feature. High values of $\chi$ (red) cluster in the negative SHAP region, which shows the systematic negative impact of molecules that are too large to fit in the channel. This shows that not only did the descriptor work as intended but also that the model itself is physically informed. This is further confirmed by RPD being the second most prominent descriptor, with larger values pushing the prediction toward adsorption likelihood. The parameters Ψ and AV are also shown to have a major impact. While secondary to $\chi$ and RPD, these descriptors have a major influence as they tell how much the molecule can move once inside the pore network. Certain molecule-based descriptors such as the first principal


https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314

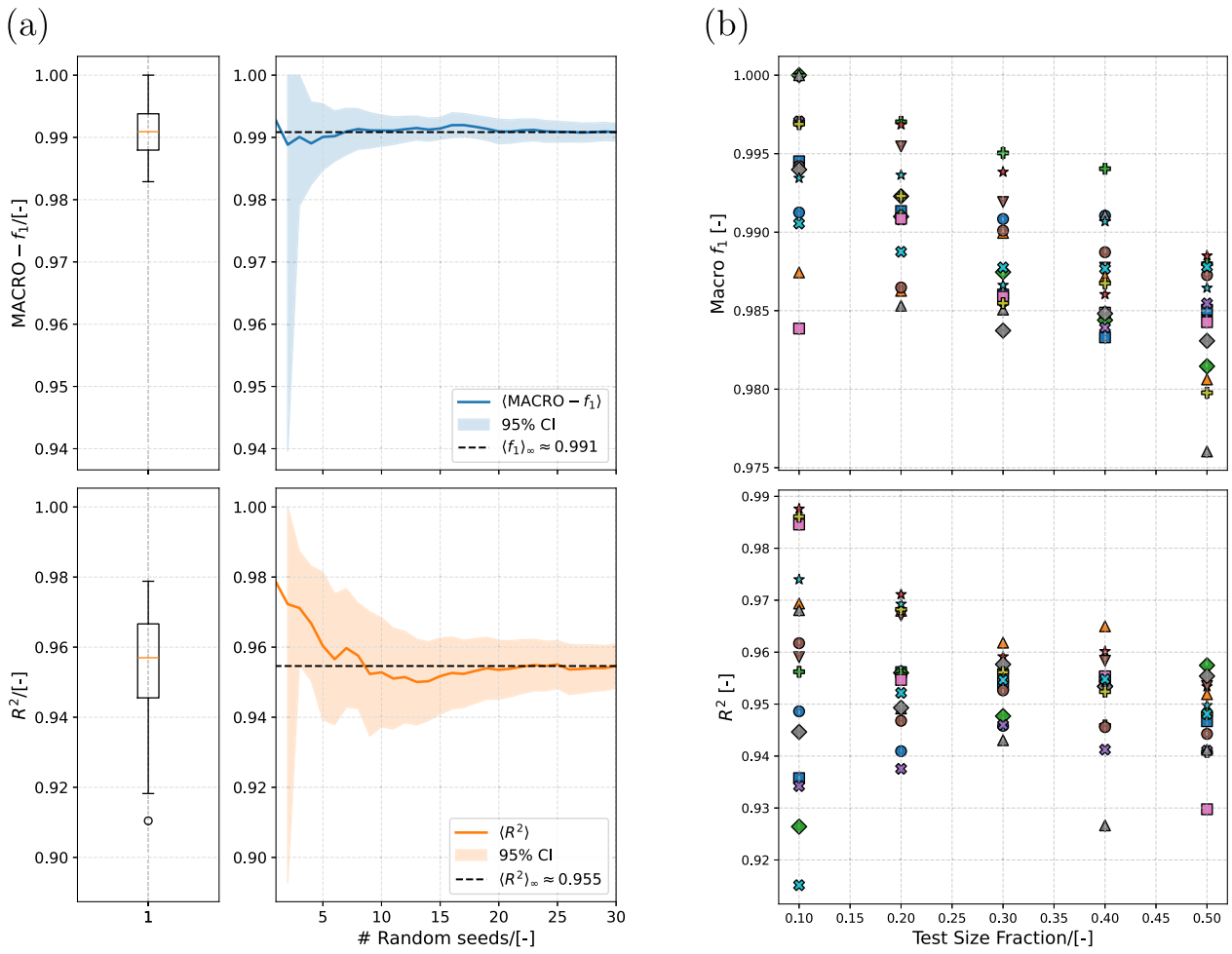



**Figure 8.** (a) Box and whisker plots as well as moving average across random seeds of both the classifier MACRO-$f_1$ (defined by eq 9) and regressor $R^2$ (defined in eq 10). The classifier performance is not affected by the changes in random seed, remaining above 0.99 for the MACRO-$f_1$. TabPFN dips slightly initially but stabilizes after a few iterations, showing the robustness of both elements of the cascade. (b) Values of the classifier MACRO-$f_1$ and regressor $R^2$ across multiple random seeds. Despite the GBT classifier decreasing in performance as the test size gets larger, the performance is still very high. However, the performance of the regressor remains acceptable but gets less consistent as the test set size is made smaller, which may indicate overfitting.

mass moment of inertia PMI1 and side-to-total carbon count ratio $R_{side}$ also influence the prediction as each serves as a proxy for the molecule's bulkiness. However, the influence of these descriptors is smaller than that of the zeolite-based descriptors as steric constraints imposed by the channels dominate the adsorption process. The remaining descriptors follow the same logic of either representing the steric constraint imposed by the channel (such as LCD, AV, NASA, $TD_{10}$ and AccV) or the overall size and bulkiness of the molecule (such as $D_{mol}^{min}$, $D_{mol}^{max}$, $\omega$, $n_{C2}$, $\frac{PMI1}{PMI2}$, and $\frac{PMI2}{PMI3}$). Another observation is that as descriptors become less important (and as distributions narrow), their influence becomes less monotonic and interpretable. This hints that the most important descriptors are used early in the prediction process and impose clear-cut boundaries. Less important descriptors are thus used for more localized splits of the data.

The SHAP distribution for the regressor featured in Figure 7b shows that the leading descriptors are the molecule volume, $V_{mol}$ and the ratio between the available zeolite volume and molecule volume, $\Psi$, meaning that the fundamental aspect behind the prediction is the volume available for a given molecule. This constitutes another steric constraint encapsulated by $\Psi$ that proves that the regressor as well is physics-informed. Notably, $\Psi$ shows a context dependence: more compact alkanes in larger channels tend to have higher maximum loadings. A noteworthy element is the presence of all three principal mass moments of inertia, the acentric factor, both minimum and maximum diameters, as well as $R_{side}$, meaning that the shape of the molecule also contributes to the final result. This proves shape-based selectivity is also taken into account to an extent. Indeed, the spreads indicate that larger, bulkier, and less spherical molecules will also have lower maximum loadings. Structural constraints given by the accessible surface area, as well as topological and framework densities, play a role, albeit minor compared to the previous two. A last remark that can be made is about how the number of side groups and their nature are being used by the models. In the classifier, it can be observed that the side-to-main chain carbon count $R_{side}$ has some influence over the final prediction, but other than the number of ethane groups, no side-group counts are present. Moreover, the regressor stage completely disregards the number of side groups but still uses the total carbon count and $R_{side}$ for its prediction. This shows that when computing the maximum uptake, the nature of the side groups has little impact. This can be traced to the choice of


https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314

descriptors: since molecular volume is already used, it implicitly accounts for the side groups. As such, the latter is used mostly in more localized splits.

### 3.5. Model Robustness

A critical aspect of the model to assess is its robustness, that is, its resilience against changes in the descriptor set or randomness in the data. In this work, three tests are carried out to assess how performance changes as essential training parameters are changed. Each component of the cascade is evaluated using a single representative metric: the classifier is assessed with MACRO-$f_1$ score and the regressor via $R^2$. For this test, TabPFN is used as the regressor, as it proved to be the best performer, and interpretability is not an issue in this context. The raw data for the study are present in the Supporting Information.

A first robustness assessment revolves around changing the random seed used to split the data set as well as initialize both the model and the Bayesian optimizer. This assesses how resilient the model is against initialization and random splits in the data. This can also give a reasonable idea of what the expected performance of the model would be across multiple runs and whether the reported metrics are representative or overly optimistic. To carry out this test, 30 random seeds are generated via a random integer generator and fed to the model.

The box-and-whiskers plots in the left column of Figure 8 show that despite an outlier on the regressor end (from random seed `210`), both models of the cascade present minimal variation. The random-seed average of the MACRO-$f_1$ score is noted to remain steady for all random seeds, with an asymptotic value of $\langle \text{MACRO-}f_1 \rangle_\infty = 0.991$. This can be attributed to the Bayesian optimization step when training the model, as it ensures that the best hyperparameters are used for the model. While the random-seed average $R^2$ of the TabPFN regressor initially dips during early iterations, it quickly stabilizes to a final value of $\langle R^2 \rangle_\infty = 0.955$. This hints that although some data splits may lead to more favorable performance on the test set, the overall expected performance is close to what was initially observed in Figure 6d. This indicates that the cascade model is robust against data splits and optimization initialization. A second assessment changes the size of the testing set. Doing so provides insight into whether the performance of the model is affected by the availability of training data. To obtain a more representative result, the average result across 15 random seeds is used. The results are shown in Figure 8b. Split sizes range from 10% to 50% of all data available at 10% increments. Despite the decrease in the MACRO-$f_1$ score as the test set grows, performance is kept high. This indicates that the classifier itself is robust against data set sizes. Furthermore, the low variance in values across the random seeds further confirms the insights from Figure 8a, showing that randomness has very limited effect on performance. The regressor is observed to have more variance across results, with the largest variance for a test data set size of 10% of all data. Since the largest variances are noted for the smaller test size fractions across random seeds, this hints that the model may operate in a high-variance regime consistent with overfitting, where the training data is very well fitted but performance on the test set is not maintained. Since TabPFN is a transformer-based model with high representational capacity, this variance reflects sensitivity to training-set composition rather than instability of the learning procedure.

### 3.6. Machine Learning versus Deep Learning

The cascade model used previously is made up of classical ML models. Increasing focus has been shifted toward using neural networks as one can use inputs that would not be very hard to code for regular ML, albeit at the cost of lower interpretability. To this end, an Artificial Neural Network (ANN) is created as a benchmark to see whether it can provide an alternative to the current cascade model. The Artificial Neural Network used in this study is implemented using PyTorch[76] and is a fully connected feed-forward network. The architecture of the network comprises an input layer with the same dimensionality as the number of descriptors, two hidden layers with 8 neurons each, and an output layer with one neuron that predicts the maximum loading as a continuous variable. Each of the hidden layers employs ReLU activation functions. The model is trained using the Adam optimizer with a learning rate of 0.001 [-] and MSE as the loss function. Training is performed across 10000 epochs. Despite the possibility to use the CUDA toolkit on the ANN to accelerate training, it was decided to use CPU-based training to keep the comparison fair, since scikit-learn and XGBoost have very limited capability to move training to the GPU. Training was done on a Dell Precision 5690 laptop with an Intel Core Ultra 7 165H processor.

We investigate the deployability of such an alternative (i.e., the amount of time and computing power needed to train the model) when compared to the current GBT classifier + TabPFN cascade. The total training time, per-sample inference time, and model size are considered. The results are presented in Table 4. It is to be noted that the time spent on Bayesian optimization is omitted, as it is only done once and does not contribute toward model runtime performance. It can be immediately observed that the wall and CPU times for training the cascade are lower than those for the ANN, thus indicating that the former is cheaper to train. This is expected since the training procedure of the neural network involves an active step of optimization to find the ideal weights of the network, while GBC is fast to optimize and TabPFN is already pretrained.

**Table 4. Cascade and ANN Training Time, Per-Sample Inference Time, and Model Size**

| | Cascade model (GBC/TABPFN) | ANN |
|---|---|---|
| **Training wall time [s]** | 0.6479/2.5151 | 18.8565 |
| **Training CPU time [s]** | 0.6406/0.7343 | 76.0625 |
| **Per-sample inference time [s/10$^{-3}$]** | 0.0024/6.3335 | 1.9377 |
| **Model size [MB]** | 0.1187/42.1429 | 0.0039 |
| **Used RAM [MB]** | 0.0078/248.2266 | 16.4531 |

The values for the inference time are seen to vary substantially between the approaches. The GBC in the cascade model is proven to be very fast, while TabPFN is much slower due to its transformer-based nature, and the ANN has an inference time between the two. However, it is to be noted that unlike the ANN, TabPFN only has to work for samples that have nonzero adsorption rather than on all samples. As such, the cascade outperforms the ANN. It is also visible that the cascade has a larger footprint in both the RAM and disk memory. No matter the case, the values at this stage are still very modest and as such do not. This does raise the possibility of using a neural network to substitute the two-stage model.



https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314

## 4. CONCLUSIONS

In this work, an ML framework was developed to predict the maximum loading of alkane isomers in zeolites. The final model relies on the combination of a gradient-boosted tree for the classifier, determining if adsorption occurs, and a regressor stage, predicting the actual maximum loading. Both take in data primarily related to the molecule and zeolite size and volume, with several topological descriptors as well as some composite descriptors to enhance the interactions between the two. The classifier labels the cases with almost perfect performance, despite having a minor bias toward predicting no adsorption in the case of doubt. Out of all considered regressors, the TabPFN model performed the best. An interpretability study was also carried out using SHAP values. It was shown that both stages provide predictions that are in line with the physics behind adsorption in zeolites, including steric hindrance and shape-based selectivity. A robustness study has been carried out on the GBT + TabPFN cascade, which showed that the full model is robust against random data splits, but the regressor performance varies a lot with too little testing data, indicating possible overfitting. Finally, the cascade model was compared to a neural network, which showed that despite a lower training time and cost, the cascade model is heavier on RAM and disk space due to the size of TabPFN. Future research on this topic may include addressing the problem of the low-value maximum uptake regime, either by altering the data set (either by adding more samples to the full data set or by oversampling the problematic region when creating the training set) or by reviewing the cascade model (e.g., having two regressors handling smaller and larger molecules, respectively). Another possible extension to this work may be the inclusion of descriptors relating to the channel connectivity, such as ones based on network science (e.g., degree or betweenness centrality). As the data set grows larger and the descriptors more complex, it may be wise to consider transitioning to deep learning methods if the accuracy is shown to improve.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcc.5c08611.

The list of considered zeolites, the list of considered alkanes, relevant information regarding the Bayesian optimization, and the raw results from the robustness study (SI1) (PDF)

All RASPA2 simulation results; it also contains the full data set of maximum loadings that is used to train all the models; this study contains 97 zeolites and 45 hydrocarbons, and therefore, 4365 data points (SI2) (XLSX)

All Python scripts used for data set generation (ML_GENDATASET.py), training and optimizing all ML models (ML_CHAINED_MODELS.py), the robustness study (ML_ROBUSTNESS.py), and training the artificial neural network (ML_NNARCHITECTURE.py); it also contains the supporting functions used to create the data, including the script that collects the IZA data online (SF_IZAONLINE.py), the sampling method used to obtain the set of considered molecules (SF_RANDMOL.py), and the set used to generate all the data needed for the alkanes using RDKit (SF_ALKANEDATA.py) (SI3) (ZIP)

## ■ AUTHOR INFORMATION


### Corresponding Author

**Thijs J. H. Vlugt** − *Process & Energy Department, Faculty of Mechanical Engineering, Delft University of Technology, Delft 2628CB, The Netherlands;* orcid.org/0000-0003-3059-8712; Email: t.j.h.vlugt@tudelft.nl

### Authors

**Eric Johnsson** − *Flight Performance & Propulsion Department, Faculty of Aerospace Engineering, Delft University of Technology, Delft 2629HS, The Netherlands*

**Shrinjay Sharma** − *Process & Energy Department, Faculty of Mechanical Engineering, Delft University of Technology, Delft 2628CB, The Netherlands; Department of Applied Physics, Eindhoven University of Technology, Eindhoven 5600MB, The Netherlands;* orcid.org/0000-0001-8345-7433

**Arvind Gangoli Rao** − *Flight Performance & Propulsion Department, Faculty of Aerospace Engineering, Delft University of Technology, Delft 2629HS, The Netherlands;* orcid.org/0000-0002-9558-8171

**David Dubbeldam** − *Van't Hoff Institute of Molecular Sciences, University of Amsterdam, Amsterdam 1098XH, The Netherlands;* orcid.org/0000-0002-4382-1509

**Sofia Calero** − *Department of Applied Physics, Eindhoven University of Technology, Eindhoven 5600MB, The Netherlands;* orcid.org/0000-0001-9535-057X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcc.5c08611


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work was also sponsored by NWO domain Science for the use of supercomputing facilities, with financial support from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (The Netherlands Organization for Scientific Research, NWO). The authors also acknowledge the use of computational resources of the DelftBlue supercomputer, provided by the Delft High Performance Computing Center (https://www.tudelft.nl/dhpc).

## ■ REFERENCES

(1) United Nations *United Nations Framework Convention on Climate Change (UNFCCC) Report of the Conference of the Parties on its Twenty-First Session, held in Paris from 30 November to 13 December 2015*; 2015. https://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf. Accessed: 11−12−2025.

(2) Wong, F. W. M. H.; Kez, D. A.; Del Rio, D. F.; Foley, A.; Rooney, D.; Abai, M. Decarbonizing and offsetting emissions in the airline industry: Current perspectives and strategies. *Energy* **2024**, *313*, 133809.

(3) Song, G.; An, H.; Hou, Y.; Tong, H.; Liu, J.; Tang, X.; Yi, H. Review of the historical trends and decarbonization pathways of the civil aviation sector. *Renewable Sustainable Energy Rev.* **2025**, *222*, 115927.

(4) van Dyk, S.; Saddler, J. *Progress in the Commercialization of Biojet/Sustainable Aviation Fuels (SAF): Technologies, potential and challenges*. 2021. https://www.ieabioenergy.com/wp-content/uploads/2021/06/IEA-Bioenergy-Task-39-Progress-in-the-



https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314

commercialisation-of-biojet-fuels-May-2021-1.pdf. Accessed: December 1, 2025.
(5) Calderon, O. R.; Tao, L.; Abdullah, Z.; Talmadge, M.; Milbrandt, A.; Smolinski, S.; Moriarty, K.; Bhatt, A.; Zhang, Y.; Ravi, V., et al. *Sustainable Aviation Fuel State-of-Industry Report: Hydroprocessed Esters and Fatty Acids Pathway*. National Renewable Energy Laboratory, 2024. , Accessed on December 1, 2025.
(6) Prussi, M.; Lee, U.; Wang, M.; Malina, R.; Valin, H.; Taheripour, F.; Velarde, C.; Staples, M. D.; Lonza, L.; Hileman, J. I. CORSIA: The first internationally adopted approach to calculate life-cycle GHG emissions for aviation fuels. *Renewable Sustainable Energy Rev.* **2021**, *150*, 111398.
(7) European Parliament and Council of the European Union Directive *European Parliament and Council of the European Union Directive (EU) 2023/2413 of 18 October 2023 Amending Directive (EU) 2018/2001 on the Promotion of the Use of Energy from Renewable Sources (Renewable Energy Directive III)*. 2023. https://eur-lex.europa.eu/eli/dir/2023/2413/oj. Accessed on February 17, 2026.
(8) Sharma, S.; Baur, R.; Rigutto, M.; Zuidema, E.; Agarwal, U.; Calero, S.; Dubbeldam, D.; Vlugt, T. J. H. Computing entropy for Long-Chain alkanes using Linear regression: Application to hydroisomerization. *Entropy* **2024**, *26*, 1120.
(9) Han, Y.; Yuan, J.; Xing, M.; Cao, J.; Chen, Z.; Zhang, L.; Tao, Z.; Liu, Z.; Zheng, A.; Wen, X.; Yang, Y.; Li, Y. Shape selectivity of zeolite for hydroisomerization of long-chain alkanes. *New J. Chem.* **2023**, *47*, 1401−1412.
(10) Misra, P.; Alvarez-Majmutov, A.; Chen, J. Isomerization catalysts and technologies for biorefining: Opportunities for producing sustainable aviation fuels. *Fuel* **2023**, *351*, 128994.
(11) Smit, B.; Maesen, T. L. M. Molecular simulations of zeolites: Adsorption, diffusion, and shape selectivity. *Chem. Rev.* **2008**, *108*, 4125−4184.
(12) Kapteijn, F.; Moulijn, J. A.; Krishna, R. The generalized Maxwell−Stefan model for diffusion in zeolites: Sorbate molecules with different saturation loadings. *Chem. Eng. Sci.* **2000**, *55*, 2923−2930.
(13) Smit, B.; Frenkel, D. *Understanding Molecular Simulation: From algorithms to applications*; Elsevier: London Wall, London, United Kingdom, 2023.
(14) Daou, A. S. S.; Fang, H.; Boulfelfel, S. E.; Ravikovitch, P. I.; Sholl, D. S. Machine Learning and IAST-Aided High-Throughput Screening of Cationic and Silica Zeolites for Alkane Capture, Storage, and Separations. *J. Phys. Chem. C* **2024**, *128*, 6089−6105.
(15) Hypothetical Zeolites. *Hypothetical Zeolites Database*. http://www.hypotheticalzeolites.net/, Accessed 11-12-2025.
(16) Earl, D. J.; Deem, M. W. Toward a database of hypothetical zeolite structures. *Ind. Eng. Chem. Res.* **2006**, *45*, 5449−5454.
(17) Rieder, S. R.; Oliveira, M. P.; Riniker, S.; Hünenberger, P. H. Development of an open-source software for isomer enumeration. *J. Cheminform* **2023**, *15*, 10.
(18) Sharma, S.; Yang, P.; Liu, Y.; Rossi, K.; Bai, P.; Rigutto, M. S.; Zuidema, E.; Agarwal, U.; Baur, R.; Calero, S.; Dubbeldam, D.; Vlugt, T. J. H. Machine Learning-Based Predictions of Henry Coefficients for Long-Chain Alkanes in One-Dimensional Zeolites: Application to Hydroisomerization. *J. Phys. Chem. C* **2025**, *129*, 18234−18249.
(19) Langmuir, I. The adsorption of gases on plane surfaces of glass, mica and platinum. *J. Am. Chem. Soc.* **1918**, *40*, 1361−1403.
(20) Vlugt, T. J. H.; Zhu, W.; Kapteijn, F.; Moulijn, J. A.; Smit, B.; Krishna, R. Adsorption of linear and branched alkanes in the zeolite silicalite-1. *J. Am. Chem. Soc.* **1998**, *120*, 5599−5600.
(21) Vlugt, T. J. H.; Krishna, R.; Smit, B. Molecular simulations of adsorption isotherms for linear and branched alkanes and their mixtures in silicalite. *J. Phys. Chem. B* **1999**, *103*, 1102−1118.
(22) Myers, A. L.; Prausnitz, J. M. Thermodynamics of mixed−gas adsorption. *AIChE J.* **1965**, *11*, 121−127.
(23) Radke, C. J.; Prausnitz, J. M. Thermodynamics of multi−solute adsorption from dilute liquid solutions. *AIChE J.* **1972**, *18*, 761−768.
(24) Walton, K. S.; Sholl, D. S. Predicting multicomponent adsorption: 50 years of the ideal adsorbed solution theory. *AIChE J.* **2015**, *61*, 2757−2762.
(25) Krishna, R.; Van Baten, J. M. How reliable is the ideal adsorbed solution theory for the estimation of mixture separation selectivities in microporous crystalline adsorbents? *ACS Omega* **2021**, *6*, 15499−15513.
(26) Makhanya, N. P.; Kumi, M.; Mbohwa, C.; Oboirien, B. Application of machine learning in adsorption energy storage using metal organic frameworks: A review. *J. Energy Storage* **2025**, *111*, 115363.
(27) Mai, H.; Le, T. C.; Chen, D.; Winkler, D. A.; Caruso, R. A. Machine Learning in the Development of Adsorbents for Clean Energy Application and Greenhouse Gas Capture. *Adv. Sci.* **2022**, *9*, 2203899.
(28) Yang, C.; Qi, J.; Wang, A.; Zha, J.; Liu, C.; Yao, S. Application of machine learning in MOFs for gas adsorption and separation. *Mater. Res. Express.* **2023**, *10*, 122001.
(29) Altintas, C.; Altundal, O. F.; Keskin, S.; Yildirim, R. Machine Learning Meets with Metal Organic Frameworks for Gas Storage and Separation. *J. Chem. Inf. Model.* **2021**, *61*, 2131−2146.
(30) Xue, X.; Cheng, M.; Wang, S.; Chen, S.; Zhou, L.; Liu, C.; Ji, X. High-Throughput Screening of Metal−Organic Frameworks Assisted by Machine Learning: Propane/Propylene Separation. *Ind. Eng. Chem. Res.* **2023**, *62*, 1073−1084.
(31) Xiuying, L.; Chen, H.; Yuan, J.; Huang, J.; Li, X.; Yu, J. Revealing the structure−property relationship of all-silica zeolites for the carbon dioxide capture: A high throughput screening study. *Zeitschrift für Naturforschung A* **2023**, *78*, 863−873.
(32) Tatlier, M.; Munz, G.; Henninger, S. K. Relation of water adsorption capacities of zeolites with their structural properties. *Microporous Mesoporous Mater.* **2018**, *264*, 70−75.
(33) Evans, J. D.; Coudert, F.-X. Predicting the Mechanical Properties of Zeolite Frameworks by Machine Learning. *Chem. Mater.* **2017**, *29*, 7833−7839.
(34) Yu, X.; Choi, S.; Tang, D.; Medford, A. J.; Sholl, D. S. Efficient Models for Predicting Temperature-Dependent Henry's Constants and Adsorption Selectivities for Diverse Collections of Molecules in Metal−Organic Frameworks. *J. Phys. Chem. C* **2021**, *125*, 18046−18057.
(35) Li, W.; Li, W.; Busch, A.; Wang, L.; Anggara, F.; Yang, S. Machine Learning Algorithm to Predict Methane Adsorption Capacity of Coal. *Energy Fuels* **2024**, *38*, 23422−23432.
(36) Zhao, L.; Zhang, Q.; He, C.; Chen, Q.; Zhang, B. J. Quantitative Structure−Property Relationship Analysis for the Prediction of Propylene Adsorption Capacity in Pure Silicon Zeolites at Various Pressure Levels. *ACS Omega* **2022**, *7*, 33895−33907.
(37) Chakraborty, A.; Gandhi, A.; Hasan, M. F.; Venkatasubramanian, V. Discovering zeolite adsorption isotherms: A hybrid AI modeling approach. *Comput.-Aided Chem. Eng.* **2024**, *53*, 511−516.
(38) Chakraborty, A.; Gandhi, A.; Hasan, M. M.; Venkatasubramanian, V. Explainable AI modeling of zeolite adsorption isotherms. *Chem. Eng. Sci.* **2026**, *320*, 122361.
(39) Guido, S.; Mueller, A. *Introduction to machine learning with python*; O'Reilly Media, 2016; p 392.
(40) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189−1232.
(41) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, 2009.
(42) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, 2016, 785−794. .
(43) Hollmann, N.; Müller, S.; Eggensperger, K.; Hutter, F. TabPFN: A Transformer that Solves Small Tabular Classification Problems in a Second. *arXiv* **2023**.


https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314

(44) Liu, P. *Bayesian Optimization: Theory and Practice Using Python*; Apress: Berkeley, CA, 2023.
(45) Dubbeldam, D.; Calero, S.; Ellis, D. E.; Snurr, R. Q. RASPA: Molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Mol. Simul.* **2016**, *42*, 81−101.
(46) Dubbeldam, D.; Torres-Knoop, A.; Walton, K. S. On the inner workings of Monte Carlo codes. *Mol. Simul.* **2013**, *39*, 1253−1292.
(47) Martin, M. G.; Siepmann, J. I. Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes. *J. Phys. Chem. B* **1998**, *102*, 2569−2577.
(48) Martin, M. G.; Siepmann, J. I. 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* **1999**, *103*, 4508−4517.
(49) Bai, P.; Tsapatsis, M.; Siepmann, J. I. TraPPE-zeo: Transferable potentials for phase equilibria force field for All-Silica Zeolites. *J. Phys. Chem. C* **2013**, *117*, 24375−24387.
(50) Liu, B.; Smit, B.; Calero, S. Evaluation of a new force field for describing the adsorption behavior of alkanes in various pure silica zeolites. *J. Phys. Chem. B* **2006**, *110*, 20166−20171.
(51) Bingel, L. W.; Chen, A.; Agrawal, M.; Sholl, D. S. Experimentally Verified Alcohol Adsorption Isotherms in Nanoporous Materials from Literature Meta-Analysis. *J. Chem. Eng. Data* **2020**, *65*, 4970−4979.
(52) Gómez-Carracedo, M. P.; Andrade, J. M.; Calvino, M.; Fernández, E.; Prada, D.; Muniategui, S. Multivariate prediction of eight kerosene properties employing vapour-phase mid-infrared spectrometry. *Fuel* **2003**, *82*, 1913−1921.
(53) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform* **2015**, *7*, 20.
(54) Baerlocher, C.; McCusker, L. B.; Brouwer, D.; Marler, B. *Database of Zeolite Structures*, 2025. https://www.iza-structure.org/databases/. Accessed 01−09−2025.
(55) Dubbeldam, D.; Calero, S.; Vlugt, T. J. H. iRASPA GPU-accelerated visualization software for materials scientists. *Mol. Simul.* **2018**, *44*, 653−676.
(56) Baerlocher, C.; McCusker; Lynne, B.; Olson, D. H. *Atlas of Zeolite Framework Types*; Elsevier, 2007; pp. 3−11.
(57) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous Mesoporous Mater.* **2012**, *149*, 134−141.
(58) Landrum, G. *RDKit: Open-source cheminformatics*. 2025. http://www.rdkit.org,. Accessed: 01−09−2025.
(59) Torres-Knoop, A.; Poursaeidesfahani, A.; Vlugt, T. J. H.; Dubbeldam, D. Behavior of the enthalpy of adsorption in nanoporous materials close to saturation conditions. *J. Chem. Theory Comput* **2017**, *13*, 3326−3339.
(60) Vlugt, T. J. H.; Schenk, M. Influence of framework flexibility on the adsorption properties of hydrocarbons in the zeolite silicalite. *J. Phys. Chem. B* **2002**, *106*, 12757−12763.
(61) Krishna, R.; Smit, B.; Calero, S. Entropy effects during sorption of alkanes in zeolites. *Chem. Soc. Rev.* **2002**, *31*, 185−194.
(62) Sharma, S.; Rigutto, M. S.; Zuidema, E.; Agarwal, U.; Baur, R.; Dubbeldam, D.; Vlugt, T. J. H. Understanding shape selectivity effects of hydroisomerization using a reaction equilibrium model. *J. Chem. Phys.* **2024**, *160*, 214708.
(63) Schenk, M.; Calero, S.; Maesen, T. L. M.; Vlugt, T. J. H.; van Benthem, L. L.; Verbeek, M. G.; Schnell, B.; Smit, B. Shape selectivity through entropy. *J. Catal* **2003**, *214*, 88−99.
(64) Li, Z.; Constantinou, L.; Baur, R.; Dubbeldam, D.; Calero, S.; Sharma, S.; Rigutto, M.; Dey, P.; Vlugt, T. J. H. Second-order group contribution method for $T_c$, $P_c$, $\omega$, and liquid densities of linear and branched alkanes. *Mol. Phys.* **2025**, *123*, No. e2566763.
(65) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.
(66) Tarzia, A. *mol-ellipsize: Molecular size calculation based on ellipsoid fitting over conformer ensembles Software*; 2021. https://github.com/andrewtarzia/mol-ellipsize. Accessed: 31−10−2025.
(67) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning*; Springer: Cham, 2013.
(68) West, R. M. Best practice in statistics: The use of log transformation. *Ann. Clin. Biochem.* **2022**, *59*, 162−165.
(69) Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Hoo, S. B.; Schirrmeister, R. T.; Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature* **2025**, *637*, 319−326.
(70) Wu, J.; Chen, X. Y.; Zhang, H.; Xiong, L. D.; Lei, H.; Deng, S. H. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *J. Electron. Sci. Technol.* **2019**, *17*, 26−40.
(71) Liu, C.; Balasubramanian, P.; An, J.; Li, F. Machine learning prediction of ammonia nitrogen adsorption on biochar with model evaluation and optimization. *Npj Clean Water* **2025**, *8*, 13.
(72) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(73) Spearman, C. The Proof and Measurement of Association between Two Things. *Am. J. Psychol.* **1904**, *15*, 72.
(74) Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2020; https://christophm.github.io/interpretable-ml-book/. Accessed on 31−01−2025.
(75) Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. NIPS, 2017; pp 4765−4774, https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.
(76) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L., et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems 32*; NeurIPS, 2019.

https://doi.org/10.1021/acs.jpcc.5c08611
J. Phys. Chem. C 2026, 130, 4299−4314