Title

Spectroscopy-Guided Deep Learning Predicts Solid-Liquid Surface Adsorbate Properties in Unseen Solvents.



Journal

Journal of the American Chemical Society, 146(1)

Authors

Du, Wenjie
Ma, Fenfen
Zhang, Baicheng
et al.

Publication Date

2024-01-10

DOI

10.1021/jacs.3c10921

Peer reviewed

JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS



Article

# Spectroscopy-Guided Deep Learning Predicts Solid−Liquid Surface Adsorbate Properties in Unseen Solvents

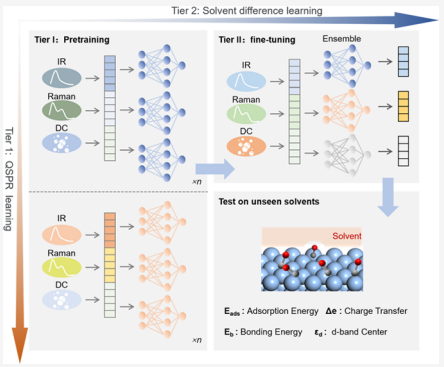
Wenjie Du,[▽] Fenfen Ma,[▽] Baicheng Zhang,[▽] Jiahui Zhang, Di Wu, Edward Sharman, Jun Jiang,* and Yang Wang*

Cite This: *J. Am. Chem. Soc.* 2024, 146, 811−823

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Accurately and rapidly acquiring the microscopic properties of a material is crucial for catalysis and electrochemistry. Characterization tools, such as spectroscopy, can be a valuable tool to infer these properties, and when combined with machine learning tools, they can theoretically achieve fast and accurate prediction results. However, on the path to practical applications, training a reliable machine learning model is faced with the challenge of uneven data distribution in a vast array of non-negligible solvent types. Herein, we employ a combination of the first-principles-based approach and data-driven model. Specifically, we utilize density functional theory (DFT) to calculate theoretical spectral data of CO−Ag adsorption in 23 different solvent systems as a data source. Subsequently, we propose a hierarchical knowledge extraction multiexpert neural network (HMNN) to bridge the knowledge gaps among different solvent systems. HMNN undergoes two training tiers: in tier I, it learns fundamental quantitative spectra−property relationships (QSPRs), and in tier II, it inherits the fundamental QSPR knowledge from previous steps through a dynamic integration of expert modules and subsequently captures the solvent differences. The results demonstrate HMNN's superiority in estimating a range of molecular adsorbate properties, with an error range of less than 0.008 eV for zero-shot predictions on unseen solvents. The findings underscore the usability, reliability, and convenience of HMNN and could pave the way for real-time access to microscopic properties by exploiting QSPR.



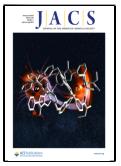

## INTRODUCTION

As known to all, surface−adsorbate microscopic interaction properties are crucial to catalysis, electrochemistry, surface molecule recognition, etc.[1−3] In fact, in material surface, adsorbate properties on solid−liquid interfaces are even more critical since the vast majority of chemical transformations occur in the liquid phase,[3] where the existing solvent can significantly influence both catalytic and reaction activities, hence leading to substantial variations in reaction rates and efficiencies.[4] Nevertheless, directly measuring these microscopic properties is still impractical and challenging. Spectroscopic tools, which are capable of measuring the dipoles carrying electronic levels and distribution information, offer the potential to establish quantitative spectra−property relationships (QSPRs) for material surface.[5]

Recently, machine and deep learning methods have sparked a paradigm shift in the analyzing and processing of spectral signal. Ultramodern studies have proposed a series of end-to-end approaches that leverage the advantages of deep learning (DL) technologies to effectively avoid the accumulation of errors in mining QSPR in the gas phase.[5,6] As a result, the utilization of DL in exploring QSPR has gained significant momentum across diverse fields, encompassing the identification mixture component,[6] characterization of microstructures,[7] nanostructured quantitative validation,[8] and so on.[9] Additionally, using a regression-based method facilitates learning an approximate formulaic mapping of QSPR, hence addressing the obstacle of imperfect and limited data.[10] However, regarding the learned end-to-end relationships, their nature of low dimension and lack of chemical explanatory factors determine the inevitably limited generalization abilities of DL-based models.

Therefore, the routine DL approach could be arduous when directly used in solvent systems due to the following two hurdles: the diversity and complexity of solvent types in real world and unbalanced distribution of available high-field data for each individual solvent in both computational and experimental perspectives. Furthermore, available high-quality data is limited to a few specific systems (i.e., gas and aqueous phases), but the majority of domain we are interested in (i.e., another phase) could differ significantly. There is a useful

Received: October 4, 2023
Revised: December 13, 2023
Accepted: December 14, 2023
Published: December 29, 2023







https://doi.org/10.1021/jacs.3c10921
*J. Am. Chem. Soc.* 2024, 146, 811−823

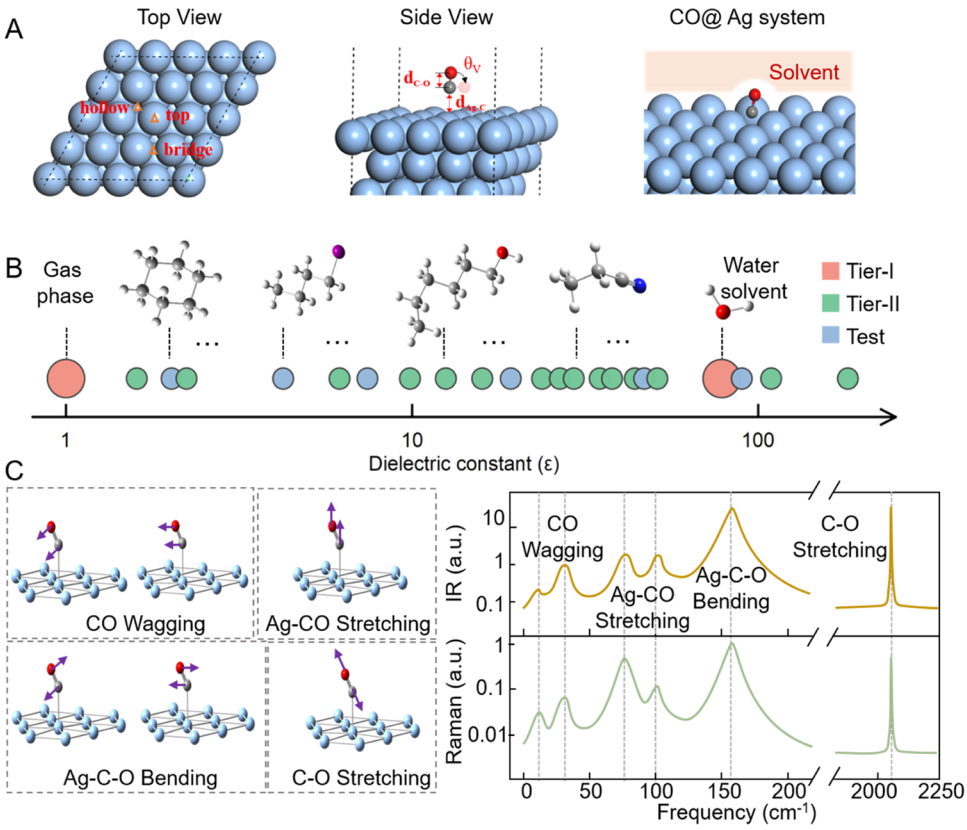



**Figure 1.** Research scope. (A) Schematic of the CO@Ag structural model construction. (B) Different solvents and corresponding dielectric constants. The dielectric constant (DC) of the gas phase is minimum and is 1.0, and the maximum dielectric constant is 181.6 in methyl formamide. The size of the circle represents the amount of data quantities. The data of water and gas phases, which contains about 2205 and 2133 records, respectively, is sufficient. And (C) six key vibrational modes (the dotted mark position; $f_1$−$f_6$; $I_1$−$I_6$; $R_1$−$R_6$) including two CO wagging modes, one Ag−CO stretching mode, two Ag−C−O bending modes, and one C−O-stretching mode. These six vibrational frequencies and IR and Raman intensities corresponding to these frequencies are chosen as the descriptors.

proposal for a model to leverage the chemical knowledge obtained from several common systems, guiding the development of a DL model for application to other, less common systems.[11] This is the essence of transfer learning, and the model could possess superior generalization capability to bridge the knowledge gaps among different solvent systems through a hierarchical transfer learning process in principle.

To fully harness the capabilities of the DL model, it is essential for it to learn as much helpful hidden knowledge from the related scenes as possible for the target field. Nowadays, individual expert module could be used to extract knowledge from decentralized data, and different experts jointly contribute to knowledge assembly through expert ensembles.[12,13] This collaborative process enables DL model to extend acquired knowledge from a narrow field to a broader scope of application.[13−15] Herein, with the enlightenment of the above-mentioned collaborative paradigm, we design and employ a hierarchical knowledge extraction multiexpert neural network (HMNN) to establish omniscient, interpretable, and robust relationships between vibrational spectral information and surface−adsorbate interaction properties on solid−liquid interfaces.

Taking Ag metal as an example, we investigate the adsorbate properties, while CO is adsorbed on Ag in 23 different solvent systems. Infrared (IR) and Raman spectral data of CO with various adsorption conformations are calculated by density functional theory (DFT),[16] which are combined with the solvent dielectric constant as input features. This model successively learns the fundamental knowledge of QSPR and explores the chemical difference knowledge among solvents from two tiers. The acquired knowledge is ensembled together as a foundation for investigating solvent-specific variations within multiple solvents. This approach is a multitasked, universal, multimodule assembly[17,18] that obtains physical insight into the system.[19,20] Based on the learned QSPR and solvent difference knowledge, this proof-of-concept protocol represents a significant milestone in predicting kinds of adsorption properties on solid−liquid interfaces with chemical precision for various adsorption conformations and previously unseen solvent systems.

## RESULTS AND DISCUSSION

A hybrid approach that combines data-driven methods with first-principle-based approaches is employed. The adsorption configurations of CO@Ag (CO adsorbed on Ag substrates) with a total number of 5703 are generated and investigated with first-principle calculations. To expand the data set's spatial distribution range, we considered variations in the distance between C and the Ag surface plane ($d_{Ag-C}$), the C−O bond length ($d_{C-O}$), and adsorption angles (vertical plane: $\theta_v$ and horizontal plane: $\theta_h$), focusing on their impact on four target properties on 23 distinct solvent systems (Table S1). On water and gas phases, the $d_{Ag-C}$ values include $d_0 - 0.2$, $d_0 - 0.15$, $d_0 - 0.1$, $d_0 - 0.0$, $d_0 - 0.1$, $d_0 - 0.2$, $d_0 - 0.3$, $d_0 - 0.4$, and $d_0 - 0.5$ Å, where $d_0$ represents the optimized structural distance.



https://doi.org/10.1021/jacs.3c10921
J. Am. Chem. Soc. 2024, 146, 811−823

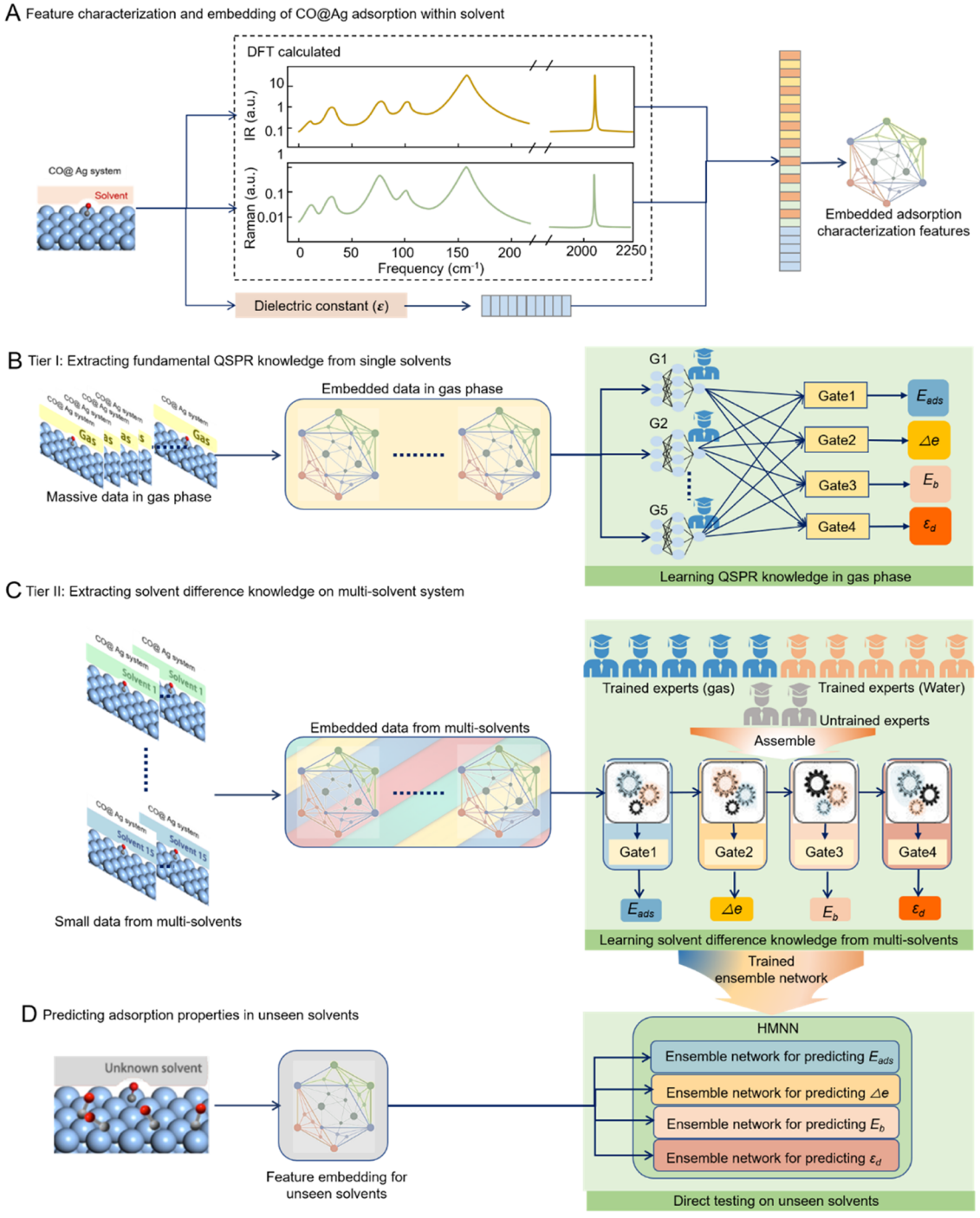



**Figure 2.** Model architecture and training strategy. (A) DFT-calculated IR and Raman spectra along with solvent dielectric constant are applied with the operations of embedding, dimension increasing, and concatenating, hence generating a unified representation of input features. (B) Tier I process. An expert network and gate network integrated multitasking learning framework for learning fundamental QSPR knowledge from the gas phase (same in the water phase). Trained experts are selected and frozen for subsequent ensemble. (C) Tier II process. Trained experts in gas and water phases are assembled with untrained experts for capturing solvent difference knowledge on a multisolvent system, which includes 15 different solvents. (D) Testing four target adsorption properties in unseen solvents.

$d_{C-O}$ is closely associated with adsorption properties, which are calculated from $d_0' - 0.04$ to $d_0' + 0.04$ Å, where $d_0'$ is the C−O bond length of the optimized configuration. The adsorption angles in the vertical plane are incremented from 0 to 80° with a small step size of 10° and in the horizontal plane are 0 and 90° (Table S2). These conformations are sampled uniformly. Regarding the multisolvent data sets including additional 15 solvents, each solvent only contains 35 relevant top-site adsorbed conformations for the training of tier II. Specifically, the $d_{Ag-C}$ is from $d_0 - 0.1$ to $d_0 + 0.3$ Å with a step size of 0.1 Å; $d_{C-O}$ is unchanged, and $\theta_v$ is only considered in 0° condition. Six additional solvents are used to validate its generalization ability in a zero-shot prediction manner. The 2-propanol (IPA), dimethyl sulfoxide (DMSO), and ethylene carbonate (EC) have the same 35 top-site adsorbed conformations in tier II, while the cyclohexane (CyH), diethyl


https://doi.org/10.1021/jacs.3c10921
J. Am. Chem. Soc. 2024, 146, 811−823

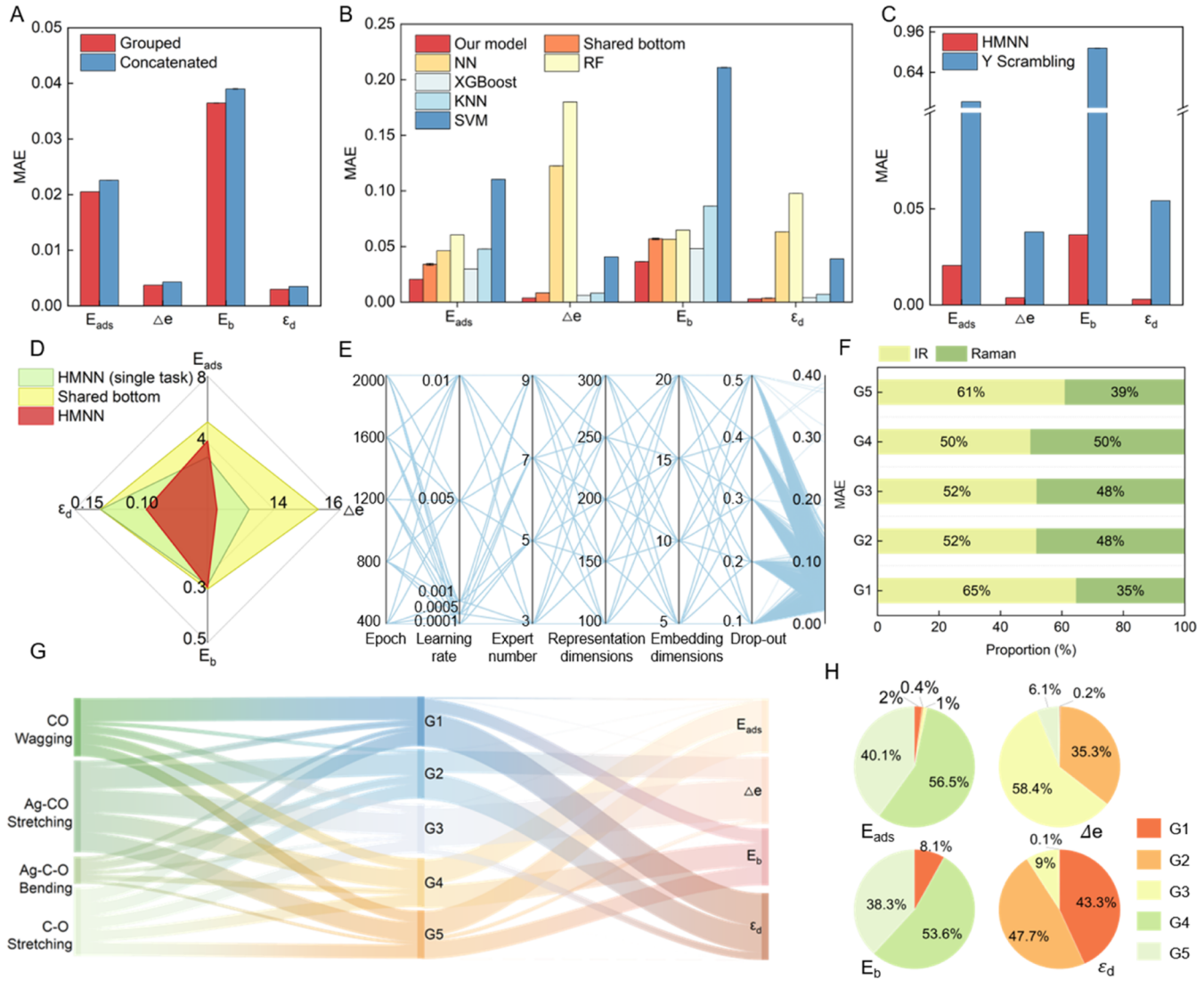



**Figure 3.** Performance of HMNN on gas data set. Performances of HMNN on gas data set. (A) The test results of different featurization methods of spectral descriptors. (B) Performances of different approaches in terms of MAE ($E_{ads}$, $E_b$, and $\varepsilon_d$ are in units of eV; $\Delta e$ in $e^-$). (C) The test results of HMNN in Y-scrambling were in terms of MAE. (D) Performance comparison of multitask and single-task modes in MAPE (%) (E) Performances of HMNN with various hyperparameter combinations; here, hyperparameters include the learning rate, expert numbers, representing dimensions and embedding dimensions, dropout, and epochs. (F) Feature importance analysis for IR and Raman spectrum. (G) Analysis of feature importance with regard to experts and expert contributions to tasks. (H) Contributions of different experts on each individual task (G1−G5: expert modules in the gas phase).

ether ($Et_2O$), and tetrahydrofuran (THF) have much richer structures, with $d_{Ag-C}$ being from $d_0 - 0.1$ to $d_0 + 0.5$ Å, a small step size of 0.1 Å, and $\theta_v$ is adjusted from 0 to 40° with intervals of 10°.

The IR and Raman spectra of aforementioned conformations are computed and documented as descriptor signals to explore QSPR, including the intensities (IR: $I_1$−$I_6$; Raman: $R_1$−$R_6$) of the IR and Raman spectrum and six corresponding vibrational frequencies ($f_1$−$f_6$) (Figure 1C). These six vibrational modes are most relevant for the CO@Ag system.[5] These features exhibit a good fit with a normal distribution (Figure S1). Regarding the frequency features, the numerical values of the six frequency features ($f_1$−$f_6$) gradually increase, with the mean of $f_6$ reaching around 2000 $cm^{-1}$. In terms of infrared and Raman spectral intensities, there is some overlap in their distributions, but $I_6$ and $R_6$ have broader distributions (Figure S2). The solvent dielectric constant is considered as an additional feature for discriminating between solutions. Eighteen vibrational spectral features and dielectric constant values are utilized as input after appropriate processing, such as direct splicing or paired input (Figure S3).

## MODEL FRAMEWORK

The HMNN hierarchical-training strategy is depicted in Figure 2. Here, the input dielectric constant is embedded by an individual network to achieve increasing dimensionality. The raw feature includes the frequency values ($f_1$−$f_6$), IR and Raman-responsive intensity values ($I_1$−$I_6$, $R_1$−$R_6$), and a dielectric constant value. We utilize 18 vibrational spectral descriptors in combination with the dielectric constant value as inputs, employing either direct splicing or paired input (Figure 2A). Two featurization approaches will be tested and selected in the next step (Figure S3).



https://doi.org/10.1021/jacs.3c10921
J. Am. Chem. Soc. 2024, 146, 811−823

The training process included 2 tiers. In tier I, two fundamental multiexpert networks are trained, respectively, on gas and water phases, as illustrated in Figure 2B. The multiexpert network contains 5 independent expert networks to learn comprehensive representations of inputted features, where the basic unit of the expert network is a two-layer fully connected network. These expert modules share the same initial input vector but operate independently without mutual interference (Figure S4). A hidden similarity control module is devised and employed in this fundamental network to ensure the differentiation and specialization of each individual expert by using the cosine similarities between every pair of expert output representations as penalty items. The learned representations are then inputted into 4 different gate networks where each gate network plays a judging role by assigning different weights to different expert networks with regard to a specific individual target task (using a SoftMax function) (Figure S5). These gate networks are initialized by utilizing the original input features as initialization parameters and then updated continually during training. Subsequently, regarding different target tasks, trained expert modules are selected, frozen, and then incorporated with some blank experts to assemble a new network for fine-tuning in tier II (Figure 2C). Fine-tuning is performed on a multisolvent system encompassing 15 different solvents to capture solvent difference. Finally, HMNN is estimated on six unseen solvents (Figure 2D).

Four molecular adsorbate properties, which are considered here as the target tasks, are set as the ultimate outputs of HMNN. In particular, they contain (1) the adsorption energy ($E_{ads}$) between the adsorbate and surface (Ag) and (2) the d-band ($\varepsilon_d$) center of the metal surface layer that are often used as indicators of catalytic activity;[21] (3) the change in charge ($\Delta e$), which is the amount of charge polarization or transfer and indicates the degree of electronic coupling between the molecule and catalyst; and finally, (4) the bond energy ($E_b$) of an adsorbate that is often used to gauge the difficulty of bond breakage. These four properties are commonly used metrics to comprehensively describe the adsorption process,[5] as determined by a combination of structural parameters and calculated by the DFT protocol (Figure S6). The calculated values align well with a normal distribution, as illustrated in Figure S7.

**Tier I: Learning Fundamental QSPR Knowledge on Single-Solvent Systems.** Five commonly used machine learning (ML) models are chosen (i.e., extreme gradient boosting (XGBoost), random forest (RF), support vector machines (SVMs), K-nearest neighbor (KNN), and fully connected neural networks (NNs)) as well as a shared-bottom multitask framework in which parameters are shared among various tasks as baselines. To evaluate the performances of all approaches to learning the fundamental QSPR knowledge on single-solvent systems, all of these baselines and HMNN are trained and evaluated on gas-phase data set where the leave-one-out cross-validation method is employed for evaluation (Tables S3 and S4). Specifically, the data set is partitioned into 10 folds, with 9 folds utilized for training and the remaining one-fold for testing. This process is repeated 30 times, and the average result is reported as the final outcome. Initially, we assessed the impact of different spectral feature input methods, as illustrated in Figure 3A (Figure S8). Clearly, paired spectral descriptors achieved a smaller prediction error (MAE result). This is likely attributed to the intentional grouping of the corresponding frequency and vibrational intensity pairs, making it easier for the model to capture the information embedded in the spectrum compared to direct concatenation spectral descriptors, leading to an improved predictive performance. Consequently, in subsequent experiments, we adopted a grouped featurization approach. Compared with other baselines, HMNN significantly outperforms all of them in terms of mean absolute error (MAE) (Figure 3B). This demonstrates that HMNN can accurately predict not only the variation tendencies but also the detailed values of the four adsorption properties (Tables S5 and S6 are for the water phase). Furthermore, we conducted Y-scrambling validation for HMNN, which is a method employed to assess whether the model's predictions are statistically significant. This validation involves shuffling the target column, replacing correct feature-target pairs with new, incorrect pairs, and then retraining the model.[22] As depicted in Figure 3C, the performance of HMNN significantly deteriorates after Y-scrambling, indicating a sharp decline in effectiveness. This suggests that the reliability of HMNN results lies in its accurate capture of QSPR relationships (detailed Y-scrambling validation results can be found in Tables S7 and S8). Besides, HMNN exhibits its absolute superiority in various tasks in a multitask manner, as shown in Figure 3D, and it averagely reduces the error by 11% in terms of MAPE on all four prediction tasks while comparing with the suboptimal approach, i.e., the shared-bottom model. Though the performances of HMNN are more superior in predicting $E_b$ and $E_{ads}$ in case it is running under a single-task mode, we discover that HMNN can produce lesser MAPE dispersion while it runs under a multitasking mode, and this demonstrates that the embedded multitasking framework of HMNN can satisfy the Pareto optimality by balancing diverse tasks (the least area in Figure 3D), likely due to the promotion of shared knowledge among tasks, facilitating learning for certain tasks despite a sacrifice in the accuracy of individual tasks.

The hyperparameters, including the learning rate, number of experts, representation dimensions, embedding dimensions, dropout rate, and number of epochs, were optimized using the random search method, as outlined in Table S5. Figure 3E illustrates the MAE results for the $E_{ads}$ predictions of the HMNN model across different hyperparameter combinations. Notably, hyperparameters such as the dropout ratio, number of experts, and learning rate exhibit a diverse range of values for each model type. Subsequently, a model is trained and tested for each hyperparameter combination connected with lines, and the corresponding results are reported. Within a series of testing scenarios, the four tasks can be affected by different hyperparameter combinations. In relative terms, smaller dropout rates or higher learning rates are correlated with increased errors. A smaller dropout rate, indicating that fewer units are dropped out during training, may lead to potential overfitting to the training data, resulting in poor performance on unseen data. Conversely, if the learning rate is too high, the model may overshoot optimal parameter values, impeding the convergence of the training process (more hyperparameter test result could be found in Figure S9).

Interestingly, we discover that the trained expert modules are naturally equipped with initial chemical senses since a subset of experts tend to learn and apply diverse spectral information to satisfy a minority of specific tasks rather than attempting to optimize performance over all tasks (Figure 3F,G). For example, in the gas phase, experts G4 and G5 have jointly acquired about 97% of all required knowledge for


https://doi.org/10.1021/jacs.3c10921
J. Am. Chem. Soc. 2024, 146, 811−823


**Figure 4.** Ensemble process of HMNN in tier II for solvent difference knowledge learning and performances of ensemble network on multisolvent data sets. 15-fold-cross-validation experiments are executed here and repeated 30 times. (A) Structure overview of the ensemble network. The top two experts from both gas and water phases for each individual task are selected and frozen for subsequent network ensemble. Here, G1−G5 and W1−W5 correspond to the trained experts, respectively, in gas and water phases. (B) Six different jury scenarios for assembling gas, water, and untrained experts. W + G + U: combination of trained water and gas experts as well as untrained experts; W + U: combination of trained water and untrained experts; G + U: combination of trained gas and untrained model's experts; only W: trained experts on water solvent; only G: trained experts on gas phase. Here, abbreviation "W, G" means trained expert modules in water solvent (W1−W5) and gas phase (G1−G5), and "U" corresponds to the untrained module. (C) Average performances of different expert combinations in different tasks on selected solvents. (D) A group cross-validation results for "W + G + U" in different tasks, and points with different colors correspond to the results on 14 selected solvents. More results are presented in Table S10, and Y-scrambling validation results are presented in Table S11. (E) Weights of different categories of experts (gas, water, and untrained) on 15 rounds of experiments. (F) T-distributed stochastic neighbor embedding (t-SNE) of various experts on multisolvent data set (for $E_{ads}$).

predicting $E_{ads}$, thus demonstrating their absolute importance to $E_{ads}$ (Figure 3H). Further, from the perspective of input features based on gradient-based analysis method[23,24] (Figure S10), an analysis reveals how the outputs of these experts are generated from intuitive spectral information. Indeed, experts G4 and G5 are significantly affected by the Ag−CO-stretching vibrational mode, which is also a major contributor to $E_{ads}$ (Figure 3G), and these two experts are also obviously affected by the C−O-stretching vibrational mode, which can be viewed as a direct contributor to $E_b$.[25] Therefore, in summary, various expert networks freely focus on distinctly different knowledge by paying different attention to different vibrational modes, while gate networks organically combine different experts to achieve knowledge synthesis on specific aspects. The collaborations between experts and gate networks including knowledge extracting, filtering, and integrating enlighten primitive chemical logics of the HMNN, thus finally supporting the accurate prediction of different tasks. For instance, in the case of learning on gas-phase data, experiments reveal that each task is significantly dominated by two top-weighted experts (i.e., the total input weights of a specific task from two top-weighted experts exceeds 90%) (similar results are also observed for experiments on water-phase data set; Figures S11 and S12). The powerful representation ability of


https://doi.org/10.1021/jacs.3c10921
J. Am. Chem. Soc. 2024, 146, 811−823

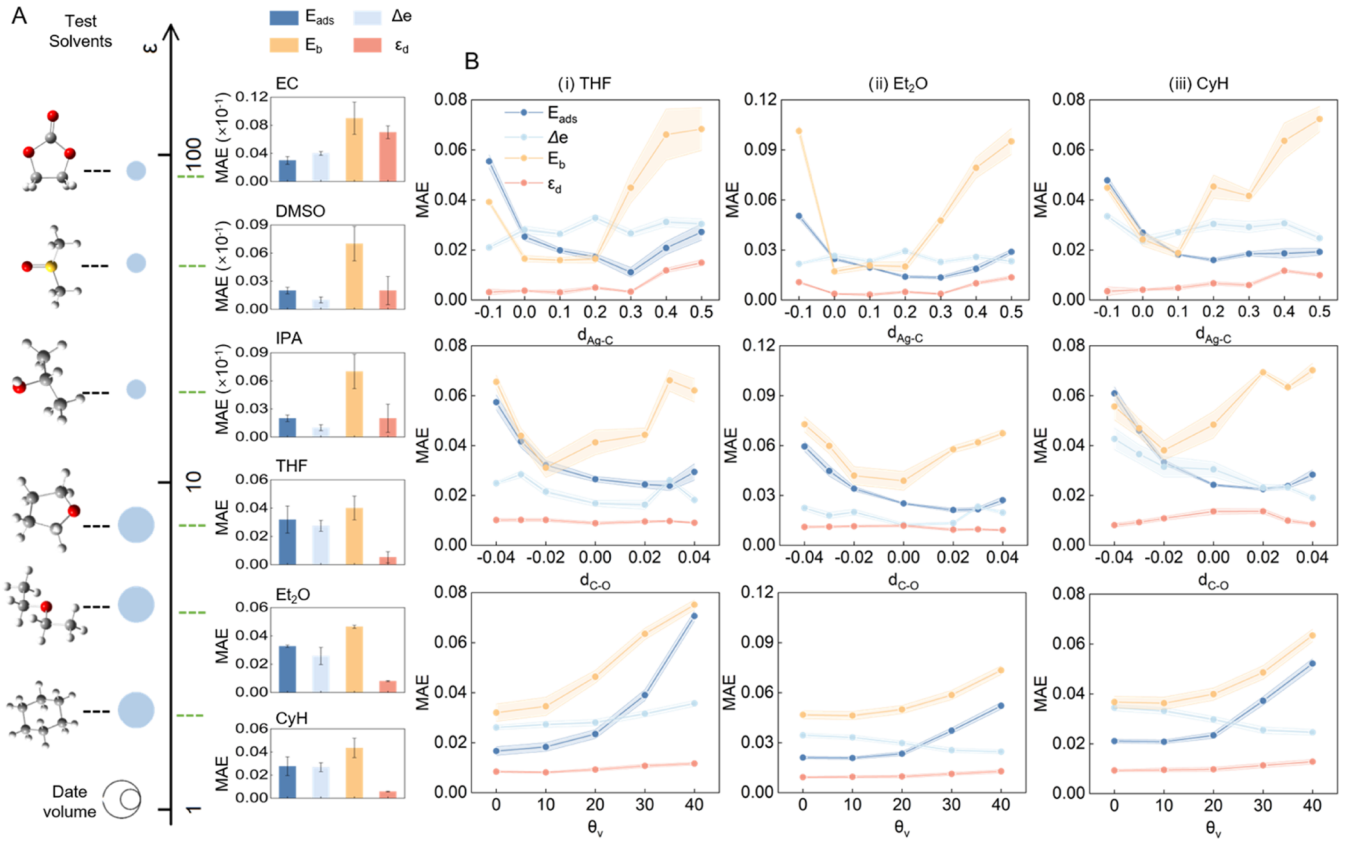



**Figure 5.** Performance of HMNN on unseen solvents. (A) The first (upper) three subfigures demonstrate the zero-shot prediction performance of HMNN on three solvents, ethylene carbonate (EC), dimethyl sulfoxide (DMSO), and 2-propanol (IPA). Each of these three solvents contains 35 data samples where each sample corresponds to the same adsorption structure that contained in the multisolvent system during tier II. The last (lower) three subfigures illustrate the performance of HMNN on the solvents of cyclohexane (CyH), diethyl ether ($Et_2O$), and tetrahydrofuran (THF) where each solvent contains 245 different adsorption structures, which are only contained in previous gas and water phases. (B) Impacts of different structural conditions on zero-shot predictions in the unseen solvents of CyH, $Et_2O$, and THF.

HMNN during tier I, which originates from its multitasking paradigm, is conducive to the learning of important spectral knowledge, hence providing a fundamental knowledge skeleton for the transfer learning process during tier II.

**Tier II: Learning Solvent Difference Knowledge on Multisolvent Systems.** Some previous studies have indicated that the performance of a trained model in a single-solvent system including gas or water phase degrades significantly once it is directly transferred to unseen systems or scenarios.[5,10] Therefore, to make sure that HMNN is omniscient to various solvents, we design a subsequent learning process (tier II) to extract solvent difference knowledge on a few-sample and multisolvent data set.[26] During tier II, for each individual task, we select the top two important experts that have already learned more than 90% of the required task knowledge in total from both gas and water phases and freeze the four selected experts (2 from gas phase and 2 from water phase) for subsequent network assembling (Figure 4A). Meanwhile, two untrained expert modules with the same structure are initialized and assembled with the four previous frozen experts, and these two new untrained experts mainly focus on extracting the unlearned knowledge from pretrained experts, hence making the ensembled network more omniscient (Figure 4A). And the multitask learning framework is also reserved and incorporated with multiple small data sets from various solvent systems in this tier to enhance the learning of solvation knowledge. Finally, the outputs of multiple decisive experts (4 experts are from the previously trained model and 2 experts are the newly added untrained experts) are fed into another round of gate networks to dynamically adjust weights for the different tasks.

Taking advantage of the flexibility and adjustability of such an ensemble learning paradigm, we have designed and tested a variety of "jury scenarios" by assembling different expert combinations (Figure 4B) and use such new combinations to replace the most quintessential combination in Figure 4A, i.e., W + G + U. The reported results of the average performance errors for different ensemble combinations undoubtedly indicate that our employed combination W + G + U significantly outperforms other alternative combinations (Figure 4C) on all four tasks. This verifies the superiority of our ensemble paradigm without challenge and simultaneously demonstrates that the ensemble network can perform better while knowledge from both gas and water phases is comprehensively involved. Concretely, the prediction performances of ensemble network (W + G + U) in cross-validation experiments demonstrate nearly perfect accuracy, and this illustrates that the ensemble network of W + G + U has captured fundamental QSPR as well as solvent difference knowledge. The weights of different categories of experts (categorized as gas, water, and untrained) in the final representations are illustrated in Figure 4E. For 15-fold-cross-validation experiments, the weights of different categories of experts are all comparable to each other, and this indicates


https://doi.org/10.1021/jacs.3c10921
J. Am. Chem. Soc. 2024, 146, 811−823

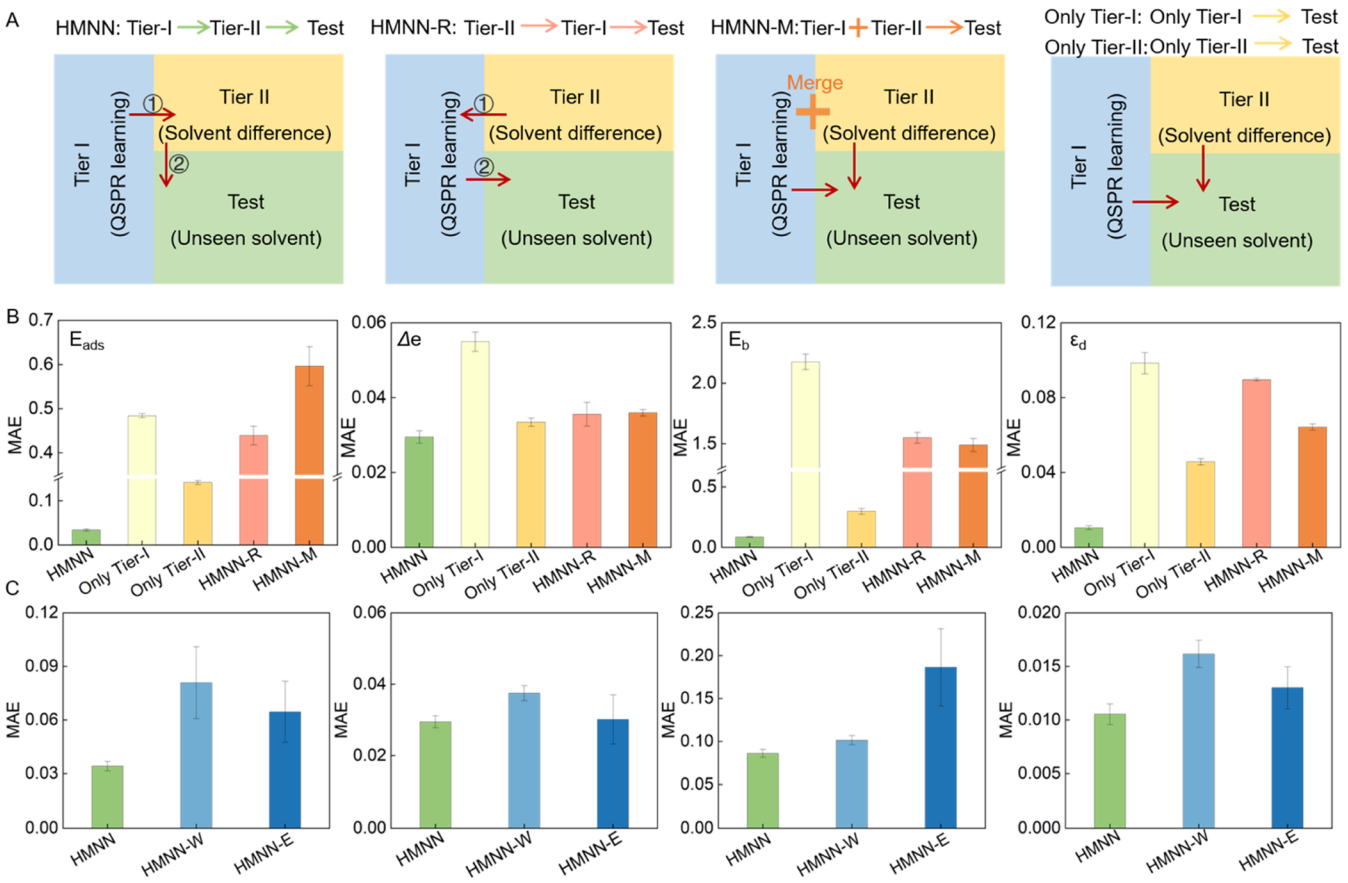



**Figure 6.** Interpretability analysis and control experiments. (A) Process diagram of 4 HMNN variants generated by using different training strategies. (B) Performance comparisons of these 4 variants and HMNN. (C) Performance comparisons of two additional variants and HMNN (taking the prediction of $E_{ads}$ on the CyH data set as an example). Variant HPMN-W is HMNN performed without a dielectric constant ($\varepsilon$), and variant HPMN-E corresponds to HMNN with an erroneous dielectric constant ($\varepsilon$).

that the newly added untrained experts play a crucial compensative role in learning new knowledge of solvent differences to make up for the shortcomings of water and air experts. Meanwhile, this figure also the embedded gate network within the ensemble network can adaptively adjust the weights of different categories of experts to make sure that the ensemble network is adaptive to different solvent systems. The T-SNE map in Figure 4F reveals that the three categories of experts are capable of learning distinct knowledge, and this also confirms the necessity of setting untrained experts.

**Predicting Adsorption Properties in Unseen Solvents with HMNN.** The hierarchical learning strategy aims at helping HMNN learn comprehensive solvent adsorption knowledge from different perspectives. During tier I, the model mostly focuses on constructing a fundamental mapping relation from spectral information to adsorption properties, i.e., QSPR. During tier II, the trained experts are assembled with untrained experts where the frozen trained experts are responsible for inheriting and maintaining basic QSPR knowledge and the newly added experts aim at learning the impacts of solvents on QSPR, i.e., solvent difference knowledge. So far, HMNN should have extensively learned and integrated the hidden knowledge that is contained in dielectric constant and spectral data. To investigate the potential of HMNN on unknown solvent systems, we conducted a series of zero-shot test experiments.

We first conduct a series of experiments on three previously unseen solvent systems, i.e., isopropyl alcohol (IPA), dimethyl sulfoxide (DMSO), and ethylene carbonate (EC). These solvents are commonly used in various applications such as the synthesis of pharmaceuticals and plastics. For all of these three solvent systems, we evaluate the average performances of all 15 rounds of experiments, and the results verify the superior and impressiveness of HMNN by achieving the chemical accuracy for all tested unseen solvents (the upper part of Figure 5A).

On the other hand, we notice that, for all previously tested solvent systems, the reactant adsorption structures of catalyst surfaces are the same as the structures in the small data of multisolvent system during tier II. This naturally brings us a question: is HMNN capable of predicting adsorption properties for an unseen solvent system, which has different adsorption structures from the structures contained in the small data of multisolvent system? This question also reflects another issue: has the knowledge learned during two different tiers been really and deeply integrated? To investigate this issue, we employ HMNN on three additional unknown solvent systems (cyclohexane (CyH), diethyl ether ($Et_2O$), and tetrahydrofuran (THF)) with 245 different adsorption structures for each individual solvent system (Table S2). As shown in the lower part of Figure 5A, the MAE prediction errors for $E_{ads}$, $\Delta e$, $E_b$, and $\varepsilon_d$ are all lower than 0.05 (units of $E_{ads}$, $E_b$, and $\varepsilon_d$ are in eV and of $\Delta e$ are in $e^-$), indicating that HMNN can stably achieve zero-shot predictions of these four adsorption properties at the level of chemical accuracy. We further analyze the MAE results of HMNN for different structural conditions (with different $d_{Ag-C}$, $d_{C-O}$, and $\theta_V$;


https://doi.org/10.1021/jacs.3c10921
J. Am. Chem. Soc. 2024, 146, 811−823

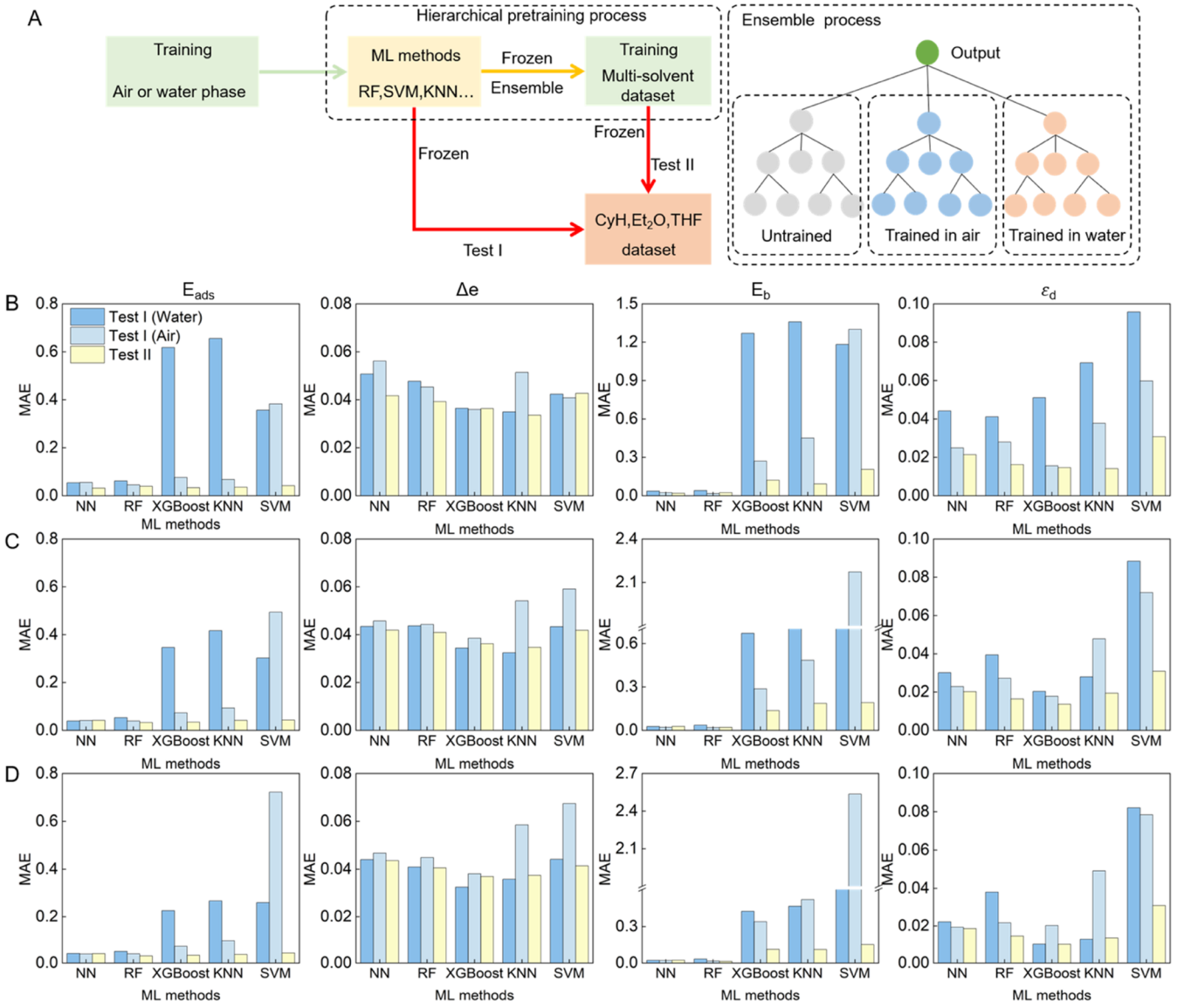



**Figure 7.** Hierarchical pretraining process for various ML methods and direct test results of various models for three distinct scenarios. (A) Diagram illustrating the workflow process. Tests on (B) CyH, (C) $Et_2O$, and (D) THF data sets. Test I: freezing the pretrained model (originating from air or water phases) and conducting a direct test. Test II: the pretrained models from air and water phases are frozen, ensembled, and fine-tuned using multisolvent data sets before subsequent testing.

Figure 5B). Actually, as can be directly observed, the prediction results are significantly impacted by several imbalanced adsorption configurations where $d_{Ag-C}$, $d_{C-O}$, or $\theta_V$ deviates from the equilibrium conformation, respectively (Figure 5B). However, for most conformations and most tasks, HMNN can still achieve chemical accuracy. This also demonstrates that, in principle, HMNN could achieve chemical precision level property predictions for various adsorption conformations in unknown solvents with only limited spectral information and dielectric constant.

**Various Training Strategy Validations and Portability Analysis of the Framework.** In this section, we delve into the factors contributing to HMNN's superior performance and explore the feasibility of transferring this framework to other ML methods. To investigate this, we first conduct a series of controlled experiments to estimate the effectiveness of our hierarchical knowledge extraction, and four new variants are generated based on HMNN (Figure 6A): (1) only tier I, which undergoes only tier I training period before being directly used to the final test; (2) only tier II, which is similar to the variant of only tier I by only employing tier II learning during its training process; (3) HMNN-R, wherein the data used in the previous two tiers is exchanged, tier I only trains 5 experts based on an integrated data set of all 15 solvents, and the top two experts with regard to a specific task are selected, frozen, and assembled with two untrained experts for subsequent training on a gas and water merged data set; and (4) HMNN-M, a HMNN variant without hierarchical training where all data sets are merged together for one integrated training period.

Obviously, the results demonstrate the importance and necessity of employing hierarchical knowledge extraction (Figure 6B). Notably, there are significant performance gaps between HMNN and the one-tier networks (only tier I and only tier II). Regarding only tier I, its performance is mostly the worst, and this indicates the importance of extracting and


https://doi.org/10.1021/jacs.3c10921
J. Am. Chem. Soc. 2024, 146, 811−823

involving solvent difference knowledge, and training only on gas and water data sets will eventually lead to the poor generalization ability of variant only tier I. However, the variant of only tier II outperforms the variant of HMNN-M. Considering that the data that used in only tier II is just a subset of the data used in HMNN-M, it is a puzzling issue that why the increase of data eventually leads to the decrease of performance? The reason may contain two aspects: (1) the extreme data disequilibrium among different solvents within the mixed data set used in HMNN-M may eventually lead to the increasing of learning bias. We think that an integrated model cannot effectively capture QSPR from such a complex and imbalanced data set; and (2) the multisolvent data set used in only tier II contains partial but insufficient useful information with regard to downstream tasks, and such information can be effectively captured by an integrated model and is of varying importance to target variables. Interestingly, this phenomenon exactly verifies the effectiveness of our hierarchical knowledge extraction. Regarding HMNN-R, even though the data sets used in two knowledge extraction tiers are exchanged, it still shows its superiority to only stage-I, and the reason maybe it can still partially capture fundamental QSPR as well as solvent difference knowledge due to its hierarchical learning nature. In summary, we can learn two things from this part: (1) solvent difference knowledge, which uses dielectric constant ($\varepsilon$) as an indication, is of great significance to downstream tasks and (2) a hierarchical knowledge extraction framework is crucial for HMNN to comprehensively extract knowledge from different dimensionalities and perspectives.

To further investigate the impacts of dielectric constant and its implicated chemical knowledge, we then generated and tested two new variants of HMNN, HMNN-W and HMNN-E, by, respectively, removing its input of dielectric constant or employing a wrong one. The evaluation results are reported in Figure 6C. As demonstrated, the performances of HMNN-E are better than those of HMNN-W in most tasks, and a possible explanation is that a randomly chosen dielectric constant may also enable the distinguishing among various solvents to a certain extent, hence providing some predictive information. This series of experiments cross-validate that, as an indicator, the dielectric constant is of great significance on distinguishing solvent difference and simultaneously emphasize the importance of employing a correct dielectric constant as well.

Additionally, we conducted portability validation experiments with various ML models. Specifically, we initially froze these pretrained ML models on water and air phases and directly tested them on unseen solvents CyH, $Et_2O$, and THF. Subsequently, the same hierarchical pretraining strategies are employed for a second round of testing to compare the prediction errors between the two instances (Figure 7A).

Certainly, several machine learning models, such as random forest and boosting algorithms, exhibit remarkable performance in single-solvent systems, particularly in water and air phases (Figure 3B), where ample training data is available. However, their performance noticeably degrades when directly applied to unseen systems, as illustrated in Figure 7 (in a zero-shot manner). We ensembled the pretrained model and applied the same hierarchical pretraining strategy on these models (Figure 7A). It is evident that, following hierarchical pretraining, the testing results of these models showed a significant improvement, accompanied by a notable reduction in the MAE metric. This enhancement is particularly pronounced for the SVM model, indicating a substantial improvement in its performance. This suggests that HMNN is also a versatile and effective conceptual framework by allowing multiple models to learn and integrate knowledge from decentralized and related domains and then fine-tuned with few data to enhance its generalization, thereby ensuring predictive capabilities in unseen scenarios.

## CONCLUSIONS

In summary, ML emerges as a valuable tool for QSPR, and we extended its applicability by incorporating considerations for solvent systems. The challenge stemming from data imbalances as well as the absence or scarcity of data among solvent systems has limited the development of ML models. Here, we leverage multiple experts to take advantage of knowledge learned from available data in related systems. Subsequently, through a hierarchical pretraining approach, we achieve a superior zero-shot prediction performance on unseen solvent systems with various adsorption structures. This implies that we can "borrow" knowledge from related systems by using a hierarchical knowledge extraction framework for the target domain. More importantly, this framework is portable, flexible, and versatile. It can be effectively combined with other ML algorithms to enhance the model's generalization ability in the target domain. Although the current experimental validation is based on DFT theoretical calculations of spectral data, the application of this proof-of-principle work, in theory, should be able to extend beyond solvation and holds great potential for practical applications. It could be viewed as a universal tool by holding the promise of addressing data imbalance challenges in areas such as solute rejection in solvent nanofiltration,[27] catalytic activity predictions for nanozymes,[28] and electrocatalyst design,[29] among others.

However, it is worth noting that high-quality data still forms the core of data-driven models.[30,31] Even though we can employ state-of-the-art techniques and models to mitigate their data requirements and enhance their out-of-distribution generalization, the scarcity or absence of data remains a bottleneck in the wider application of machine learning models. On the other hand, the current HMNN framework necessitates the training of multiple expert networks to capture domain-specific knowledge and involves multistage fine-tuning processes. This also implies that the model is larger, demanding increased computational resources and a longer training time.

## MATERIALS AND METHODS

The model consists of three modules: an embedding module, an expert and gate network integrated learning module, and an output module. The raw inputted features only contain IR and Raman spectra and dielectric constants, which are mapped and concatenated to embedding vectors related to chemical information by the embedding module. Then, in the learning module, expert networks are in charge of generating a corresponding representation for each data sample, and gate networks are responsible for generating a specific task representation for each task using both experts' outputs and raw input features. Finally, a task-specific layer, i.e., the output module, extracts information from task representations and completes the prediction. We describe the implementation of each individual module in detail in the following subsections.

**Data Generation.** Structural optimizations and electronic descriptors were computed at the density functional theory (DFT) level implemented by the Vienna Ab Initio Simulation Package


https://doi.org/10.1021/jacs.3c10921
J. Am. Chem. Soc. 2024, 146, 811−823

(VASP) with the frozen-core all-electron projector augmented wave (PAW) model[32] and Perdew–Burke–Ernzerhof (PBE) functions.[33] A kinetic energy cutoff of 400 eV was used for the plane-wave expansion of the electronic wave function, and the convergence criteria of force and energy were set to 0.01 eV Å$^{-1}$ and $10^{-5}$ eV, respectively. A Gaussian smearing of 0.1 eV was applied for optimizations and a $k$-point grid with a 3 × 3 × 1 $\gamma$ centered mesh for sampling the first Brillouin zone. The slab models of metal surfaces contain 4 layers of a 4 × 4 × 1 supercell, in which the bottom 2 layers were fixed at the bulk crystal geometry during structural optimization. To avoid artificial interactions between layers, a vacuum spacing of ∼20 Å was applied. The effect of solvent was taken into account with an implicit solvation model.[16]

Adsorption energy is calculated by $E_{ads} = E_{sur} + E_{mol} - E_{sur-mol}$, in which $E_{sur}$, $E_{mol}$, and $E_{sur-mol}$ represent the energies of the pristine metal surface, the free-standing CO molecule, and the adsorbed configurations, respectively. Cluster models cut from the periodic models were applied for the spectral calculations, as depicted in Figure 1C. IR and Raman calculations were computed with Gaussian 16 using the PBEPBE functional with the 6-31+G* basis set for the main elements and the pseudo-LANL2DA basis set for Ag atoms.[34] The Raman cross section was calculated by

$$\left(\frac{d\alpha}{d\Omega}\right)_k = \frac{\pi^2}{\varepsilon_0^2}(\tilde{\nu}_{in} - \tilde{\nu}_k)^4 \frac{h}{8\pi^2 c\tilde{\nu}_k}\left(\frac{45\alpha_k'^2 + 7\gamma_k'^2}{45}\right) \frac{1}{1 - \exp(-hc\tilde{\nu}_k/k_B T)}$$

**Featurization and Embedding of Spectra and Dielectric Constant.** Generally, a spectrum is a two-dimensional curve wherein each point represents two values: frequency and responsive intensity. Herein, we formalize the points on a spectral curve as a binary array and then input the array into the embedding layer. The raw feature is an 18-dimensional vector (the frequency values, and IR and Raman-responsive intensity values of the different vibrational modes, as illustrated in Figure 1); to make better use of these data, we perform some feature processing based on chemical knowledge. Specifically, we reassemble these 18 features into 12 feature pairs ($p_i$) ($i$ is the pair number ranging from 1 to 12) by choosing a frequency and the corresponding responsive value of one spectrum. Each feature pair is used as a descriptor for passing through a two-layer fully connected network, and we eventually get 12 pairs of embedding vectors. Then, we concatenate them with a raw feature $x$ as the final embedding vector result ($e$), i.e.

$$e = x \| \mathrm{FC}(p_1) \| \mathrm{FC}(p_2) \| \cdots \| \mathrm{FC}(p_{12})$$

where the function FC corresponds to a fully connected neural network and ‖ denotes the operation of concatenation.

Each dielectric constant ($\varepsilon$) is a single value, and for a specific solvent, there exists only one corresponding dielectric constant value. So, a two-layer fully connected neural network is used here to generate a 20-dimensional vector (more details in the Supporting Information), which is then concatenated with $e$, i.e.

$$m = e \| \mathrm{FC}(\varepsilon)$$

where the function FC corresponds to a fully connected neural network and ‖ denotes the operation of concatenation.

**Expert and Gate Network Integrated Learning Module.** This module can be divided into two parts: an expert submodule and a gate submodule.35 To begin, we use 5 expert networks $f_i()$ ($i$ indexes the number of expert networks and is from 1 to 5) to learn the final embedding vector result and use this result to obtain different hidden representations. An expert network consists of a multilayer perceptron with the LeakReLU activation function and a normalization layer[35] and finally outputs representations with regard to this expert. Second, after obtaining all 5 experts' representations, gate networks $G_j$ ($j$ here indexes the number of gate networks, which also corresponds to task indexes one by one and ranges from 1 to 4) are trained to learn how to "score the opinions of different experts", meaning gate networks can assign weights to different experts to meet the needs of different tasks. Specifically, a gate network, which is a series of linear transformations with a SoftMax, utilizes the original input features as initialization parameters and keeps updating during training. For a specific task, the gate network can be calculated by

$$G_j = \mathrm{SoftMax}[\mathrm{FC}(x)]$$

where FC is a two-layer fully connected network. Here, $G_j$ is a five-element weight parameter array, which can also be written as $G_j = [g_1^j g_2^j g_3^j g_4^j g_5^j]$, and naturally, the sum of all weights within $G_j$ satisfies

$$\sum_{i=1}^{5} g_i^j = 1$$

where $j$ is the task index, and $i$ is the expert module index. Then, the final representation with regard to the $j$th task, i.e., $t_j$, is calculated by

$$t_j = \sum_{i=1}^{5} g_i^j f_i(m)$$

Thus, by integrating gate networks with expert networks, each task is associated with an exclusive task-specific representation $t_j$ ($j$ is the number of gate networks ranging from 1 to 4 and also corresponds to task indexes one by one). To ensure the diversity among different experts, we further introduce a cosine similarity-based similarity control (SC) module. It is a penalty term, which is calculated during forward propagation and minimized by backward propagation to reduce the similarities between task-specific representations, i.e.

$$(t_j, t_k) = \frac{t_j \cdot t_k}{\| t_j \| \cdot \| t_k \|}$$

where a‖·a‖ indicates L2 normalization.

**Feature Weight Calculation.** We calculated the features' weight based on gradient-based analysis method[23,24] (Figure S10). Specifically, we computed the absolute value of its corresponding gradient and summed the absolute gradient values of features

$$C_i = \sum_{j=1}^{m} \left| \frac{\partial L}{\partial w_{ij}} \right|$$

This sum represents the overall sensitivity of the model to all of the input features.

Here, $i$ represents the input feature, $m$ is the number of nodes in the hidden layer, $j$ represents the hidden layer nodes, and $w$ signifies the weights from input features to hidden layer nodes. Consequently, the contribution $P$ of $n$ features can be calculated using the following formula

$$P = \sum_{i=1}^{n} C_i$$

**Output Module.** This module aims to decode the representations from task-specific gate networks and finally output prediction results. Each task-specific output module includes three same-stacked block submodules where each block submodule consists of one fully connected layer, one LayerNorm layer, and a LeakReLu activation function. An additional fully connected layer is employed as the final layer following the three stacked blocks to output the final result.

**Loss Function.** For a single prediction task, MSE is a good metric for evaluation; therefore, we use such a metric here. However, when evaluating multiple tasks, it is trivial to maintain the balance between the losses of different tasks since they may have different measurement scales. Therefore, we here introduce an elegant way,[36] which is also a principled way of combining multiple loss functions to simultaneously learn multiple objectives by using homoscedastic uncertainty.

**Architecture of Ensemble Network in Tier II.** The ensemble network has nearly the same architecture as the raw expert and gate network integrated learning model, except the number of experts


https://doi.org/10.1021/jacs.3c10921
J. Am. Chem. Soc. 2024, 146, 811–823

differs. Several trained experts with frozen parameters in parallel are contained in a previously trained model to involve the extraction of fundamental QSPR knowledge. Simultaneously, several untrained experts with the same structure are also integrated to extract solvent difference knowledge during subsequent training.

## ■ ASSOCIATED CONTENT

### Data Availability Statement

The HMNN code and generated data sets are available at https://github.com/invokerqwer/HMNN.

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.3c10921.

Detailed data set description, model details, and test results (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Authors

**Jun Jiang** − *Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, China; School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China;* orcid.org/0000-0002-6116-5605; Email: jiangj1@ustc.edu.cn

**Yang Wang** − *Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, China; School of Software Engineering, University of Science and Technology of China, Hefei, Anhui 230026, China; Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu 215123, China;* orcid.org/0000-0002-6079-7053; Email: angyan@ustc.edu.cn

### Authors

**Wenjie Du** − *Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, China; School of Software Engineering, University of Science and Technology of China, Hefei, Anhui 230026, China; Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu 215123, China*

**Fenfen Ma** − *Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, China; School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China; Gusu Laboratory of Materials, Suzhou, Jiangsu 215123, China*

**Baicheng Zhang** − *Key Laboratory of Precision and Intelligent Chemistry, University of Science and Technology of China, Hefei, Anhui 230026, China; School of Chemistry and Materials Science, University of Science and Technology of China, Hefei, Anhui 230026, China;* orcid.org/0000-0002-1899-028X

**Jiahui Zhang** − *School of Software Engineering, University of Science and Technology of China, Hefei, Anhui 230026, China; Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu 215123, China*

**Di Wu** − *School of Software Engineering, University of Science and Technology of China, Hefei, Anhui 230026, China; Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu 215123, China*

**Edward Sharman** − *Department of Neurology, University of California, Irvine, California 92697, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.3c10921

### Author Contributions

[▽]W.J. D., F.F.M., and B.C.Z contributed equally to this work.


### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This paper was partially supported by the Project of Stable Support for Youth Team in Basic Research Field, CAS (YSBR-005), the Natural Science Foundation of China−National Major Research Instrument Development Project (No.12227901), the Innovation Program for Quantum Science and Technology (2021ZD0303303), the National Natural Science Foundation of China (22025304, 22033007), and the GHfund C (202302038682).

## ■ REFERENCES


(1) Zhao, X.; Liu, Y. Unveiling the Active Structure of Single Nickel Atom Catalysis: Critical Roles of Charge Capacity and Hydrogen Bonding. *J. Am. Chem. Soc.* **2020**, *142*, 5773−5777.

(2) Zhao, X. H.; Liu, Y. Y. Origin of Selective Production of Hydrogen Peroxide by Electrochemical Oxygen Reduction. *J. Am. Chem. Soc.* **2021**, *143*, 9423−9428.

(3) Gould, N. S.; Li, S.; Cho, H. J.; et al. Understanding solvent effects on adsorption and protonation in porous catalysts. *Nat. Commun.* **2020**, *11*, No. 1060.

(4) Reichardt, C. Solvents and solvent effects: an introduction. *Org. Process Res. Dev.* **2007**, *11*, 105−113.

(5) Wang, X.; Jiang, S.; Hu, W.; et al. Quantitatively Determining Surface-Adsorbate Properties from Vibrational Spectroscopy with Interpretable Machine Learning. *J. Am. Chem. Soc.* **2022**, *144*, 16069−16076.

(6) Fan, X.; Ming, W.; Zeng, H.; Zhang, Z.; Lu, H. Deep learning-based component identification for the Raman spectra of mixtures. *Analyst* **2019**, *144*, 1789−1798.

(7) Lansford, J. L.; Vlachos, D. G. Infrared spectroscopy data- and physics-driven machine learning for characterizing surface microstructure of complex materials. *Nat. Commun.* **2020**, *11*, No. 1513.

(8) Verma, S.; Chugh, S.; Ghosh, S.; Rahman, B. M. A. A comprehensive deep learning method for empirical spectral prediction and its quantitative validation of nano-structured dimers. *Sci. Rep.* **2023**, *13*, No. 1129.

(9) Giambagli, L.; Buffoni, L.; Carletti, T.; Nocentini, W.; Fanelli, D. Machine learning in spectral domain. *Nat. Commun.* **2021**, *12*, No. 1330.

(10) Chong, Y.; Huo, Y.; Jiang, S.; et al. Machine learning of spectra-property relationship for imperfect and small chemistry data. *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120*, No. e2220789120.

(11) Moret, M.; Friedrich, L.; Grisoni, F.; Merk, D.; Schneider, G. Generative molecular design in low data regimes. *Nat. Mach. Intell.* **2020**, *2*, 171−180.

(12) Qi, T.; Wu, F.; Wu, C.; et al. Differentially private knowledge transfer for federated learning. *Nat. Commun.* **2023**, *14*, No. 3785.

(13) Patel, D.; Wong, G. GPT-4 Architecture, Infrastructure, Training Dataset, Costs, Vision, MoE. Demystifying GPT-4: The Engineering Tradeoffs That Led OpenAI to Their Architecture. *SemiAnalysis* **2023**, *10*, 1−17.

(14) Open AI. GPT-4 Technical Report, arXiv preprint arXiv:2303.08774v3. arXiv.org e-Print archive, 2023 https://doi.org/10.48550/arXiv.2303.08774.

(15) Zhou, W. Y.; Yang, Y.; Yu, C.; et al. Ensembled deep learning model outperforms human experts in diagnosing biliary atresia from sonographic gallbladder images. *Nat. Commun.* **2021**, *12*, No. 1259.
(16) Mathew, K.; Sundararaman, R.; Letchworth-Weaver, K.; Arias, T. A.; Hennig, R. G. Implicit solvation model for density-functional study of nanocrystal surfaces and reaction pathways. *J. Chem. Phys.* **2014**, *140*, No. 084106.
(17) Cao, Y.; Geddes, T. A.; Yang, J. Y. H.; Yang, P. Y. Ensemble deep learning in bioinformatics. *Nat. Mach. Intell.* **2020**, *2*, 500−508.
(18) Ganaie, M. A.; Hu, M.; Malik, A. K.; Tanveer, M.; Suganthan, P. N. Ensemble deep learning: A review. *Eng. Appl. Artif. Intell.* **2022**, *115*, No. 105151.
(19) Tkatchenko, A. Machine learning for chemical discovery. *Nat. Commun.* **2020**, *11*, No. 4125.
(20) Loh, C.; Christensen, T.; Dangovski, R.; Kim, S.; Soljacic, M. Surrogate- and invariance-boosted contrastive learning for data-scarce applications in science. *Nat. Commun.* **2022**, *13*, No. 4223.
(21) Liu, D.; Zhao, Y.; Wu, C.; et al. Triggering electronic coupling between neighboring hetero-diatomic metal sites promotes hydrogen evolution reaction kinetics. *Nano Energy* **2022**, *98*, No. 107296.
(22) Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and its variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345−2357.
(23) Zhang, Q.-s.; Zhu, S.-C. Visual interpretability for deep learning: a survey. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 27−39.
(24) Selvaraju, R. R. et al. *Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization*, Proceedings of the IEEE International Conference on Computer Vision; IEEE, 2017.
(25) Wang, X. J.; Ye, S.; Hu, W.; et al. Electric Dipole Descriptor for Machine Learning Prediction of Catalyst Surface-Molecular Adsorbate Interactions. *J. Am. Chem. Soc.* **2020**, *142*, 7737−7743.
(26) Ma, J.; Fong, S. H.; Luo, Y.; et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat. Cancer* **2021**, *2*, 233−244.
(27) Ignacz, G.; Szekely, G. Deep learning meets quantitative structure−activity relationship (QSAR) for leveraging structure-based prediction of solute rejection in organic solvent nanofiltration. *J. Membr. Sci.* **2022**, *646*, No. 120268.
(28) Wei, Y.; Wu, J.; Wu, Y.; et al. Prediction and design of nanozymes using explainable machine learning. *Adv. Mater.* **2022**, *34*, No. 2201736.
(29) Liu, J.; Luo, W.; Wang, L.; et al. Toward excellence of electrocatalyst design by emerging descriptor-oriented machine learning. *Adv. Funct. Mater.* **2022**, *32*, No. 2110748.
(30) Vermeire, F. H.; Green, W. H. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chem. Eng. J.* **2021**, *418*, No. 129307.
(31) Guo, Y.; Li, G.; Mabuchi, T.; et al. Prediction of nanoscale thermal transport and adsorption of liquid containing surfactant at solid-liquid interface via deep learning. *J. Colloid Interface Sci.* **2022**, *613*, 587−596.
(32) Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **1994**, *50*, No. 17953.
(33) Paier, J.; Hirschl, R.; Marsman, M.; Kresse, G. The Perdew−Burke−Ernzerhof exchange-correlation functional applied to the G2−1 test set using a plane-wave basis set. *J. Chem. Phys.* **2005**, *122*, No. 234102.
(34) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B. et al. et al. In *Modeling Task Relationships in Multi-Task Learning with Multi-Gate Mixture-of-Experts*, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018; pp 1930−1939.
(35) Ba, J. L.; Kiros, J. R.; Hinton, G. E. Layer Normalization, arXiv:1607.06450. arXiv.org e-Print archive, 2016. https://doi.org/10.48550/arXiv.1607.06450.
(36) Kendall, A.; Gal, Y.; Cipolla, R. In *Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; IEEE, 2018; pp 7482−7491.


https://doi.org/10.1021/jacs.3c10921
J. Am. Chem. Soc. 2024, 146, 811−823