JCTC Journal of Chemical Theory and Computation

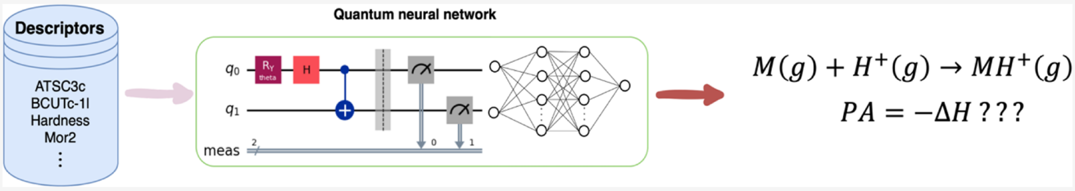
Integrating Machine Learning and Quantum Circuits for Proton Affinity Predictions

Hongni Jin and Kenneth M. Merz, Jr*

Cite This: *J. Chem. Theory Comput.* 2025, 21, 2235−2243

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: A key step in interpreting gas-phase ion mobility coupled with mass spectrometry (IM-MS) data for unknown structure prediction involves identifying the most favorable protonated structure. In the gas phase, the site of protonation is determined using proton affinity (PA) measurements. Currently, mass spectrometry and *ab initio* computation methods are widely used to evaluate PA; however, both methods are resource-intensive and time-consuming. Therefore, there is a critical need for efficient methods to estimate PA, enabling the rapid identification of the most favorable protonation site in complex organic molecules with multiple proton binding sites. In this work, we developed a fast and accurate method for PA prediction by using multiple descriptors in combination with machine learning (ML) models. Using a comprehensive set of 186 descriptors, our model demonstrated strong predictive performance, with an R^2 of 0.96 and a MAE of 2.47 kcal/mol, comparable to experimental uncertainty. Furthermore, we designed quantum circuits as feature encoders for a classical neural network. To evaluate the effectiveness of this hybrid quantum-classical model, we compared its performance with traditional ML models using a reduced feature set derived from the full set. A correlation analysis showed that the quantum-encoded representations have a stronger positive correlation with the target values than the original features do. As a result, the hybrid model outperformed its classical counterpart and achieved consistent performance comparable to traditional ML models with the same reduced feature set on both a noiseless simulator and real quantum hardware, highlighting the potential of quantum machine learning for accurate and efficient PA predictions.

■ INTRODUCTION

Identification of metabolites is important because it provides useful information in order to understand metabolic diseases, facilitate the implementation of precision medicine and helps elucidate the pathways of metabolic networks.[1−3] A widely used, effective and complementary method (e.g., to MS and NMR) to aid in the identification of the structure of unknown compounds is ion mobility coupled to mass spectrometry (IM-MS).[4,5] In positive mode IM-MS studies, a proton is added to a molecule, resulting in the formation of the $[M + H]^+$ adduct ion at the most basic site in the molecule. When multiple basic sites are available, protonation may occur at different sites or at several sites simultaneously to generate ions with multiple monoprotonation sites. The protonation site has a profound effect on the three-dimensional structure of the ion in the gas-phase and thus influences the observed collisional cross section (CCS) value. Recently an *in silico* CCS calculation workflow has been developed in our group. In this workflow,[6] the first step is to list all possibly favored protonated states using, largely, chemical intuition. As expected, given a molecule, different protonated states generate different CCS values, however, only one of them generally matches the experimental CCS value. Hence, quickly locating the most favored protonation site can save much effort in CCS calculation resulting in the acceleration of the identification of an unknown compound. In the gas phase, the proton affinity (PA) facilitates the identification of the most favored protonation site of a molecule. PA is defined as the negative enthalpy change of the protonation process of a molecule in the gas-phase where a proton is attached to a gaseous species at a specific location,

$$M(g) + H^+(g) \rightarrow MH^+(g); \quad PA = -\Delta H$$

PAs can be determined both experimentally and theoretically. In 1971, the first gas-phase ion−molecule equilibrium studies were reported.[7−9] Typically, the experimental determi-

Received: November 26, 2024
Revised: January 31, 2025
Accepted: January 31, 2025
Published: February 17, 2025

nation of PA is based on mass spectrometry, including photoionization mass spectrometry,[10] Fourier Transform Ion Cyclotron Resonance (FT-ICR) mass spectrumetry,[11] high pressure mass spectrometry (HPMS),[12] etc. With these methods, significant effort has been expended to measure the PA of various compounds and these results were collected and tabulated in 1998.[13] However, the measurement of PA is still challenging since most compounds are nonvolatile and easily thermolabile which limits experimentation in the gas-phase.

Along with experimental techniques, multiple *ab initio* computational methods are also available for the study of acid/base properties. One obvious advantage of theoretical methods is that the absolute PA can be directly derived for most molecules.[13,14] The most widely used computational methods are the Gn series (G1, G2, G3, G4),[15−19] GnMP2 (n = 2 ,3, 4)[20−22] and Weizmann1 (W1, W1BD).[23] Among them, G4 and W1 are the most accurate methods with an error of ±1 kcal/mol. However, both methods are time-consuming, and can only be applied to small to medium-sized molecules. For large systems, these theoretical methods become too computationally intensive.[24] Considering both experimental and computational methods have explicit draw backs which limit the investigation of PA, it is prudent to explore other options for PA estimation.

Quantitative structure−property relationships (QSPRs) have a long history in chemistry but have enjoyed a renaissance due to the rapid advances in machine learning (ML) and artificial intelligence (AI) bolstered by significant advances in computing resources. ML has been widely used in chemistry to correlate chemical structures with their physical-chemical properties,[25−27] predict reaction outcomes[28,29] as well as to facilitate the *de novo* design of new molecules and materials.[30−33] ML leverages statistics techniques to interpret the QSPRs accurately and efficiently. In this work, we first present a simple, fast and accurate approach to predict PAs of molecules with diverse structures. We use a full set of 186 molecular descriptors including 2*D*/3D physicochemical, quantum and fingerprint types as input features for ML models to bypass limitations of current experimental and computational methods with accuracies comparable to experiment.

In addition, we explore the potential advantages of quantum machine learning (QML). As an innovative and rapidly emerging technique, QML takes advantage of the unique properties of quantum bits, or qubits, to achieve computational capabilities that far surpass those of classical computing in certain domains. Unlike classical bits, which exist in a state of either 0 or 1, qubits leverage the principles of quantum mechanics, such as superposition, entanglement, and quantum interference to exist in multiple states simultaneously, exponentially increasing the computational space and enabling QML to solve complex problems more efficiently than classical approaches. The manipulation of qubits is facilitated by quantum operations performed using quantum gates, which act as the building blocks of quantum circuits. These gates are analogous to classical logic gates but operate under the laws of quantum mechanics. Single-qubit gates, such as the Pauli gates (X, Y, and Z) and the Hadamard gate, allow for the manipulation of individual qubits. For example, the Hadamard gate is commonly used to place qubits in a superposition state with equal probabilities. In addition, multiple-qubit gates, such as the Controlled NOT (CNOT) gate and the SWAP gate, enable interaction and entanglement, which creates correlations between qubits to process information in ways that are impossible in classical systems. Similar to classical digital circuits, quantum circuits are constructed by chronologically ordering quantum gates. These circuits perform quantum operations that manipulate the quantum states of qubits through a series of transformations. At the end of the quantum computation, a measurement operation is applied to collapse the qubits' quantum states into classical states, allowing classical information to be extracted and interpreted.

A typical workflow of QML for classical data is to preprocess classical data into quantum states, design parameterized quantum circuits,[34] and postprocess the quantum computation results to return them in classical format. In the preprocessing stage, classical data is encoded into quantum states using methods such as amplitude encoding or angle encoding, allowing it to be processed by quantum systems. Next, parameterized quantum circuits are designed, consisting of quantum gates with tunable parameters optimized during training to solve tasks like classification or regression. These quantum circuits exploit quantum properties such as superposition and entanglement to explore complex solution spaces. After computation, measurement collapses the quantum states into classical outcomes, which are then postprocessed to produce interpretable results, such as predictions or classifications. This workflow bridges classical data with quantum computational power, paving the way for enhanced machine learning capabilities. In the current noisy intermediate-scale quantum (NISQ) era, quantum computing suffers from the effect of noise that affects the ability to scale available qubits. However, promising results have demonstrated the potential applications of various QML algorithms in chemistry, such as drug toxicity prediction,[35,36] molecule design,[37,38] energy estimation.[39−41] Herein, we propose a hybrid quantum neural network (QNN) for PA prediction. The parameterized quantum circuit serves as a feature encoder, embedding input features into a quantum-enhanced representation that is subsequently processed by a classical neural network (NN). Using the same reduced feature set derived from the full feature set, this hybrid QNN outperforms its classical NN counterparts and some traditional ML methods, demonstrating the superior expressive power of quantum circuits for feature embedding.

■ METHOD

Data Curation. The data set used in this study was collected from the NIST WebBook database.[13] The simplified molecular-input line-entry system (SMILES) was used as the individual identification for each entry. The SMILES string of each molecule was retrieved from its CAS ID in the PubChem database, and molecules whose CAS ID were not included in PubChem were removed. Since we focus on the PAs of "organic" metabolites, molecules including other elements except N, P, O, S were also discarded. The data set was further curated following a protocol developed by Fourches et al.,[42] including the removal of radical species; keeping the average value of stereoisomers if the difference of these values was less than 1 kcal/mol, otherwise keeping both values; and standardization of chemical structures using RDKit.[43] Finally, 1185 compounds were left with proton affinities ranging from 150 to 260 kcal/mol.

Descriptors Calculation. *Physicochemical Descriptors.* Before calculating the descriptors, all molecules were optimized using the MMFF94 force field. In total, we computed 1826 descriptors using Mordred.[44] The descriptors with missing values, near-zero-variance, or high internal correlation coefficient ≥0.9 were excluded.

Quantum-Chemical Descriptors. Seven descriptors were calculated, including the energy of the highest molecular occupied orbital ($\varepsilon_{\text{HOMO}}$), the energy of the lowest unoccupied molecular orbital ($\varepsilon_{\text{LUMO}}$), chemical potential $\left(\mu = \frac{\varepsilon_{\text{HOMO}} + \varepsilon_{\text{LUMO}}}{2}\right)$, hardness $\left(\eta = \frac{\varepsilon_{\text{LUMO}} - \varepsilon_{\text{HOMO}}}{2}\right)$, dipole moment, the most negative atomic charge obtained using the Merz−Kollmann (MK) method[45] and Charge Model 5,[46] respectively. The geometry of each molecule was optimized using the B3LYP density functional method at the 6-31G (d, p) basis set in Gaussian 16.[47]

Molecular Fingerprint. This method transforms structural information into binary vectors. For a given molecule, the composition of its fingerprint depends on whether it includes the substructure from a list of predefined structural keys. Here, we used MACCS keys, a substructure keys-based fingerprint with 167 bits.[48] The fingerprints of all molecules in this study were retrieved using RDKit.[43]

Feature Selection. To avoid redundant features, the importance of each feature was calculated using the built-in feature importance algorithm in XGBoost. To identify the optimal features, we ranked all features by importance values and progressively removed features, starting from the least important, until we observed a significant decline in the model's performance. Finally,186 descriptors were kept (Table 1) and all descriptor values were normalized to transform the mean and the standard deviation of each descriptor into zero and one, respectively.

Table 1. Number of Descriptors Used in Each Type

Type	Count	Source
Physicochemical	86 (2D-descriptor) 14 (3D-descriptor)	Mordred-1.2.0
Quantum-chemical	7	Gaussian 16
MACCS fingerprint	79	RDKit

Similarity Calculation. To evaluate the diversity of the data set, we calculated the similarity of all compounds. We used Morgan2 fingerprints generated by RDKit to evaluate the Tanimoto coefficient,[49]

$$S_{(\text{A,B})} = \frac{c}{a + b - c}$$

where c is the number of bits that overlap in both molecules A and B; a and b are the number of bits in molecules A and B, respectively.

ML ALGORITHMS

Traditional ML Methods. Several traditional ML algorithms were explored in this study, including Support vector regressor (SVR), Random Forest regressor (RFR) and extreme gradient boosting (XGBoost) and Gradient boosting decision tree (GBDT). SVR is adapted from support vector machines (SVMs) which is originally used for classification. SVR aims at finding a hyperplane in high dimensions to fit the data so that the total error cost is minimized. The other three algorithms are all ensemble models, *i.e.*, a couple of base models (weak learners) combined together to form a strong learner, thus improving the accuracy of the model. RFR is a regression algorithm that leverages the contribution of multiple decision tress. Each node in the decision tree predicts the output based on a random subset of features. The individual output of each decision tree is averaged to generate the final output. Both XGBoost and GBDT are a tree boosting system, where each decision tree is created in sequential form to correct the errors made by previously trained trees.

Hybrid QNN. To maximally use features with minimal qubits, we adapted the patch method[50] proposed for image generation. The patch method uses several identical quantum circuits with different parameters as a subgenerator, and the same input is shared among these subgenerators but since each quantum circuit has different parameters, the output of each subgenerator is unique and by patching these outputs from subgenerators, a complete image is generated. The patch method can greatly alleviate the need for large-scale qubits because each circuit with only several qubits can be run sequentially on the same quantum device or in parallel across multiple devices. Meanwhile, the patch method avoids the entanglement of qubits at a long distance, thus narrowing down the noise effects. To match the topology of quantum devices, the virtual qubits are first transpiled to physical qubits on hardware, and if qubits are far away from each other, swap gates are necessary to connect both qubits, which dramatically increases the depth of quantum circuits and the number of gates, leading to increased noise. But since in the patch method each circuit only has a few qubits, it is easy to find physical qubits with full connectivity, hence extra gates can be avoided. In this work, we used Élivágar[51] to generate efficient and noise-resistant circuits as a subencoder for PA prediction. As an efficient Quantum Circuit Search (QCS) method, Élivágar considers the noise impact from the device topology on circuit-mapping, thus allowing early rejection of low-fidelity circuits. With the generated circuits, we explored the performance of hybrid QNN in terms of the number of features, the number of qubits, the number of parameterized gates, and the number of circuits. For a given feature set, the features were scaled to [0, π], and evenly split into subsets. The hybrid QNN model includes several structurally identical subencoders, each of which takes one feature subset for angle embedding. The expectation values after measurement are then concatenated as the input of a classical neural network. One hybrid model example is shown in Figure 1. The quantum circuits were trained utilizing TorchQuantum.[52] Subsequent performance evaluations of the trained models were conducted on test data set using both noiseless simulators and the IBM-Cleveland quantum hardware through Qiskit.[53] The neural network in all hybrid QNN models follows a consistent framework which includes three fully connected layers with the following dimensions: (input_features, input_features/2), (input_features/2, input_features/4), (input_features/4, 1), where the input_features are the measurement value of the quantum circuit. A sigmoid activation function was used in the last layer and the target PA values were scaled to [0,1] to make sure that the outputs of the model are in the correct target range. To evaluate the performance of each model, all outputs were scaled back to the original range.

RESULTS AND DISCUSSION

Diversity of the Data Set. The diversity of the chemical structures in the data set is an important metric to assess the accuracy and range of applicability of a model. With a diverse data set, models usually have higher generalization, *i.e.*, they are better to forecast unseen data. The global structure diversity

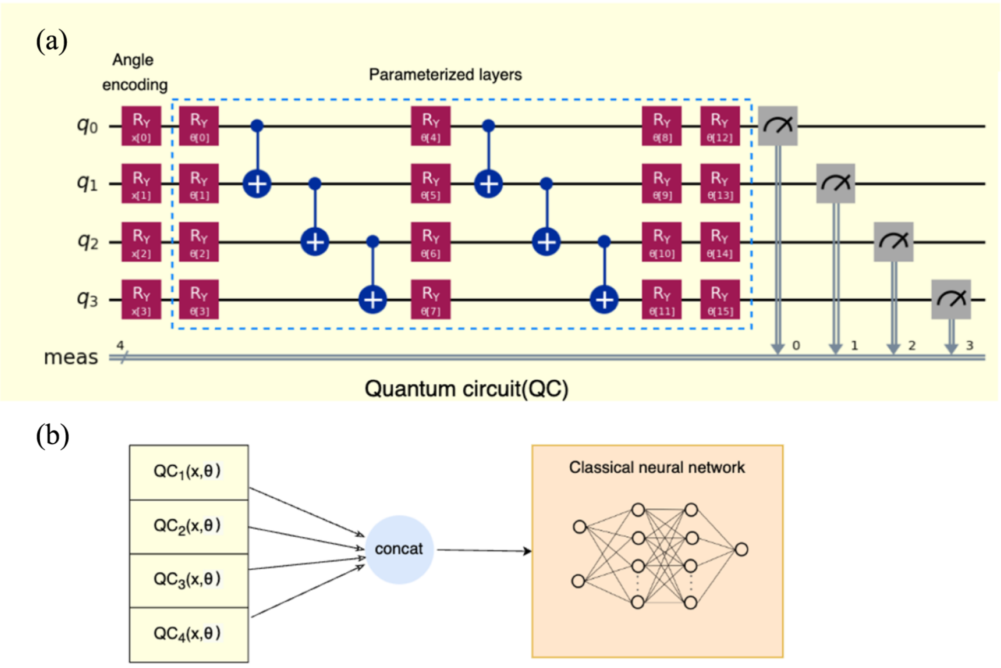


Figure 1. Hybrid QNN model with 4 qubits. (a) Parameterized circuit as subencoder for feature embedding of a neural network. The angle encoding strategy is used to transform classical input feature data into quantum Hilbert space. The parameterized layers have trainable parameters which are optimized during the training process. The quantum circuit has 16 trainable parameters $\theta \in R$ with 4 input features $x \in R$. Each qubit is measured to obtain classical information. (b) Four structurally identical subencoders with different input features x and trainable parameters θ are concatenated as the feature encoder which is then fed into a classical neural network.

was evaluated using Tanimoto coefficients. The calculated mean Morgan2 similarity is 0.067 with a standard deviation of 0.07. The Butina clustering algorithm[54] was used to group molecules with similar structures to the same cluster at a similarity level of 0.7. And 1013 singletons were generated plus 58 clusters, among which the largest one has only 8 molecules. These results suggest that the overall data set covers a diverse structural space.

Model Performance. We tested the whole data set via 5-fold cross validation. Twenty independent iterations were performed to get an unbiased evaluation on each model. All hyperparameters were tuned by a grid search method. The metrics to evaluate the models include coefficient of determination (R^2), mean absolute error (MAE) and the root-mean- square-error (RMSE). The results are shown in Table 2. SVR and GBDT performed better than RFR and XGBoost. We then combined both models together using an ensemble algorithm (Voting Regressor) implemented in scikit-learn with the weights of 1.5:1. The results of this combined model are shown in Table 2. Not surprisingly, the Voting Regressor has the best performance, since it balances out the individual weaknesses of each model. The MAE value of the Voting Regressor method is close to the estimated experimental uncertainty (~2 kcal/mol)[13] of the whole data set which suggests with selected descriptors, the Voting Regressor ensemble method is able to predict the PA at experimental accuracy but much more efficiently than other methods. Figure 2 plots a typical 5-fold cross validation of the ML predicted PAs (Pred_PA) vs the experimental PAs (Exp_PA).

Table 2. Overall Performance of Each Model[a]

ML model	R^2	MAE	RMSE
SVR	0.946 ± 0.002	2.665 ± 0.049	3.983 ± 0.092
RFR	0.930 ± 0.002	3.272 ± 0.038	4.689 ± 0.052
GBDT	0.948 ± 0.002	2.822 ± 0.038	4.041 ± 0.060
XGBoost	0.941 ± 0.001	3.030 ± 0.025	4.322 ± 0.051
Voting Regressor	0.958 ± 0.001	2.467 ± 0.039	3.633 ± 0.051

[a] The statistics is reported in the format as "mean ± standard deviation" for the 5-fold cross validation with 20 iterations. The error unit is kcal/mol.

Hybrid QNN. In this section, the entire data set was randomly split into training and test sets at a ratio of 8:2, respectively. We tested the performance of the hybrid QNN with 16, 32, and 64 features selected from Table 1. For the number of qubits, we tested 4, 8, and 10 qubits. In terms of the number of trainable gates in the quantum circuit, we manually specified it in the set of {12, 20, 40, 64}.

The results are listed in Table 3. The overall trend is as expected, *i.e.*, with more features, the accuracy of the model improves and the increase of the subencoder also yields lower errors. For example, we tested the performance of the hybrid model with 16 and 32 features using 2 or 4 subencoders. With 16 features, the MAE of 4 subencoders is 5.88 kcal/mol which is lower than that of 2 subencoders (5.93 kcal/mol). Similarly, given 32 features, 4 subencoders clearly outperform 2 subencoders, decreasing the MAE from 5.21 to 4.77 kcal/mol. The total number of trainable parameters for quantum circuits in both models is identical, but since 4 subencoders give rise to more features as inputs of the classical neural network, the hybrid model with 4 subencoders is more expressive than that with 2 subencoders. The increase of qubits also improves the accuracy of the hybrid model. Given 32 features, the hybrid model of 4 subencoders, each of which has 8 qubits decreases the MAE from 4.77 kcal/mol to 4.03 kcal/mol, compared with the equivalent subencoder which has only

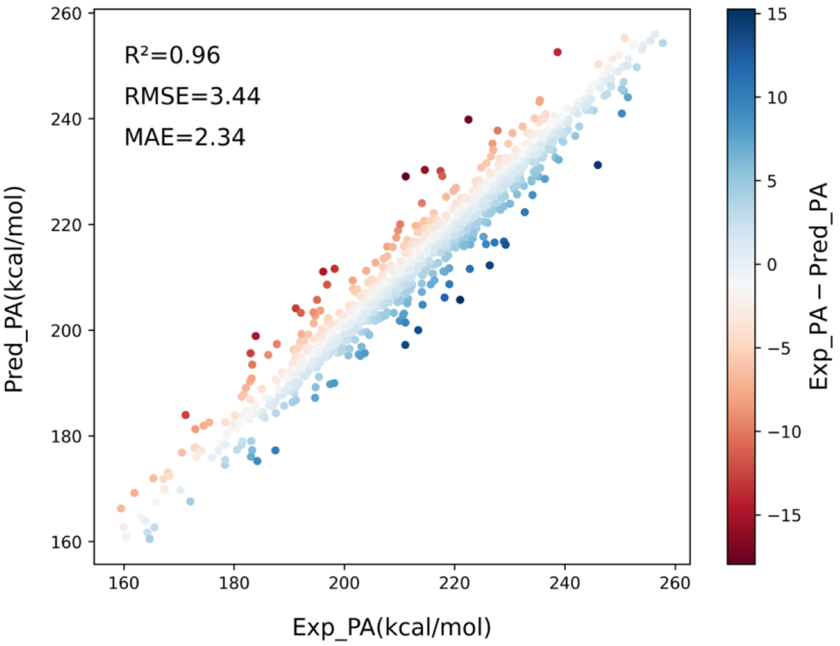


Figure 2. Overall accuracy of the Voting Regressor method.

Table 3. Performance of Hybrid QNN Model with Regard to the Number of Qubits, Features, Subencoders and Parameterized Gates[a]

Qubit	T_{Features}	$N_{\text{Subencoder}}$	N_{Features}	N_{Params}	T_{Params}	Dep/T_{gates}	R^2	MAE	RMSE
4	16	4	4	12	225	8/24	0.81	5.88	7.97
	16	2	8	20	153	16/43	0.81	5.93	7.76
	32	4	8	20	257	15/42	0.86	4.77	6.72
	32	2	16	40	129	35/83	0.85	5.21	7.04
	64	4	16	40	337	32/83	0.91	3.94	5.38
8	32	4	8	20	753	5/39	0.91	4.03	5.66
	64	4	16	20	753	11/57	0.92	3.52	5.18
	64	4	16	40	833	18/86	0.92	3.50	4.99
	64	4	16	60	913	28/122	0.93	3.42	4.74
10	64	4	16	40	1201	16/87	0.93	3.31	4.70
	64	4	16	64	1297	25/130	0.94	3.29	4.59

[a] T_{Features}: the total number of features encoded in the hybrid model; $N_{\text{Subencoder}}$: the number of subencoder in the hybrid model; N_{Features}: the number of input features per quantum circuit (subencoder); N_{Params}: the number of parameterized gates per quantum circuit (subencoder); T_{Params}: the total number of trainable parameters in the hybrid model, *i.e.*, the sum of the parameters in the quantum circuit and neural network. Dep/T_{gates}: the depth and the number of gates per quantum circuit. All models were run on a noiseless simulator. The error unit is kcal/mol.

Table 4. Performance of Various Models with 16, 32, and 64 Features[a]

Features	16			32			64		
Metrics	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE
SVR	0.78	7.20	8.51	0.85	5.76	6.94	0.87	5.43	6.62
RFR	0.83	5.62	7.53	0.88	4.46	6.22	0.92	3.58	5.22
GBDT	**0.85**	**5.27**	**6.97**	**0.91**	**3.99**	**5.54**	0.93	**3.20**	4.61
XGBoost	0.81	6.12	7.95	0.89	4.20	5.82	0.93	3.40	4.91
NN	0.78	6.41	8.40	0.88	4.78	6.15	0.92	3.63	5.06
Hybrid QNN	0.81	5.88	7.97	0.91	4.03	5.66	**0.94** (0.89)	3.29 (3.63)	**4.59** (5.24)

[a] For the hybrid QNN model, the best results of each feature ensemble in Table 3 are reported. The parameterized model with 64 features was also run on IBM-Cleveland hardware, and the results are given in parentheses. Best results of each feature ensemble are shown in bold.

4 qubits. The same trend was also observed for 64 features. The hybrid model of 4 subencoders, each of which has 4 qubits with 40 trainable parameters achieves a MAE of 3.94 kcal/mol, whereas the equivalent subencoder with 8 qubits results in a MAE of 3.52 kcal/mol. And the analogous subencoder with 10 qubits further decreases the MAE to 3.31 kcal/mol. In all cases, the subencoder modules have the same number of trainable parameters, but with more qubits, the interactions between qubits are more complex, which increases the expressivity of quantum circuits, leading to more expressive feature embeddings. Finally, we investigated the performance of the hybrid model in terms of the number of trainable gates in a

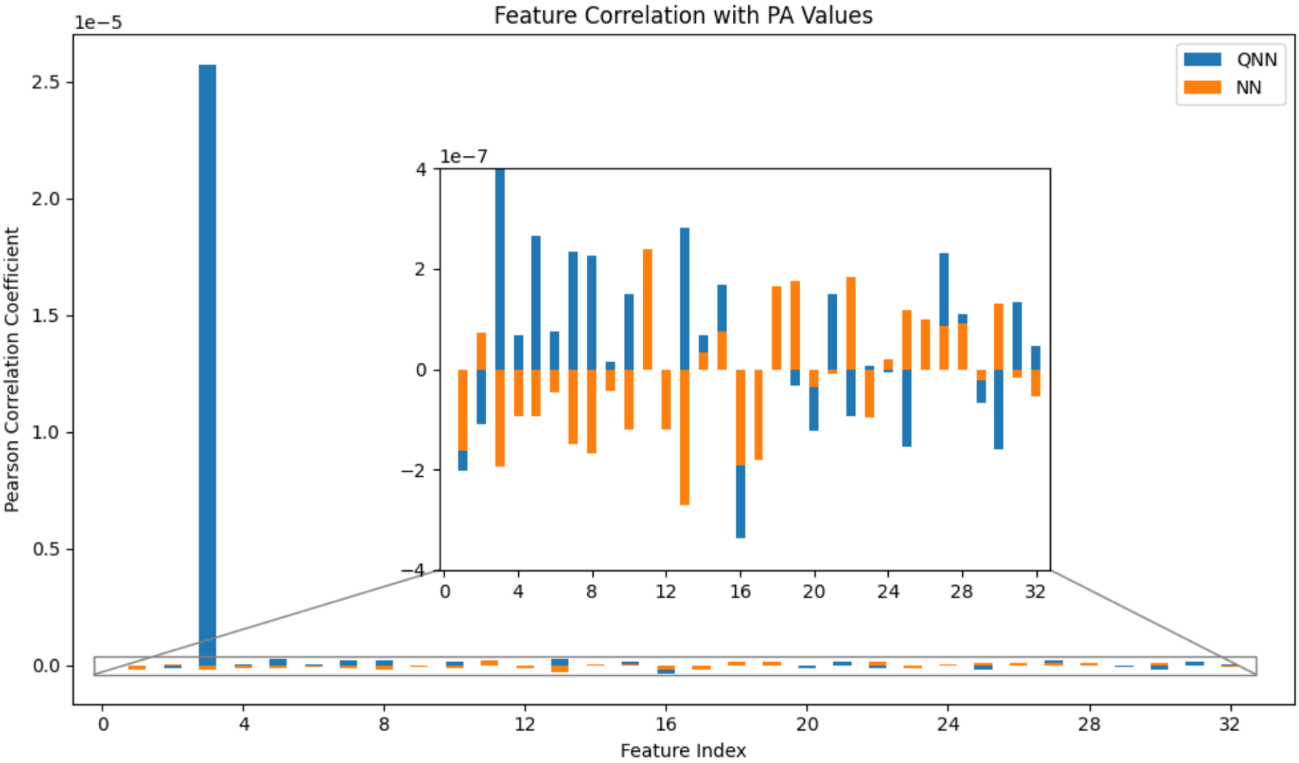


Figure 3. Correlation coefficients between 32 features and PA values in QNN and NN. The zoom-in window displays the actual range of correlation coefficients for all features except the feature at index = 3 in QNN.

single subencoder. Given 8 qubits, we generated quantum circuit with 20, 40, and 60 trainable gates as a single subencoder and the MAE decreases from 3.52 to 3.50 kcal/mol to 3.42 kcal/mol, respectively. Similarly, with 10 qubits, the subencoder with 64 trainable gates outperforms the subencoder with 40 trainable parameters, decreasing the MAE from 3.31 to 3.29 kcal/mol.

In addition, we compared the hybrid QNN model with traditional ML methods as well as a classical NN. The results are shown in Table 4. To ensure a fair comparison, the classical NN adopts the same framework as that used in the hybrid model. In all three cases, the hybrid model consistently outperforms its classical NN counterpart, demonstrating the performance of quantum circuits as feature encoders. Additionally, the hybrid model requires significantly fewer parameters than the classical NN. For example, with 64 features, the classical NN utilizes 2625 trainable parameters, whereas the hybrid model achieves a reduced MAE from 3.63 to 3.29 kcal/mol with fewer than half the parameters. This efficiency indicates the potential of quantum circuits in scaling to large models which usually have millions of trainable parameters. By replacing certain linear layers in the NN with quantum circuits, the resulting lightweight models could substantially lower computational costs, enabling faster training and inference times. Moreover, in all three cases, GBDT achieves the overall best accuracy, followed by the hybrid model which however makes the greatest progress with more features. For instance, by increasing 16 features to 32 features, the hybrid model achieves a MAE reduction of 1.85 kcal/mol and a RMSE decrease of 2.31 kcal/mol, notably surpassing the improvements that other ML models could achieve. This ability to achieve the greatest progress with more features underscores the hybrid model's superior capacity to capture and process the additional information provided by expanded feature sets. While other ML models may struggle to translate increased input dimensions into significant performance gains, the hybrid model leverages quantum circuits' ability to explore high-dimensional spaces efficiently. This allows it to extract richer representations and maximize the utility of the added features. Such a capability is particularly advantageous in scenarios where complex data sets with numerous variables are involved, as the hybrid model can continue to improve its performance rather than plateau. This demonstrates its scalability and adaptability to high-dimensional tasks, where traditional models often face challenges due to the curse of dimensionality or inefficiencies in feature utilization.

Finally, we ran the parameterized circuit with 64 features three times on real hardware and reported the best result in Table 4. Current quantum computers are susceptible to noise from various sources, leading to unavoidable errors in quantum computations. To mitigate error effects, we used the dynamical decoupling strategy. The hybrid model implemented on hardware yields a MAE of 3.63 kcal/mol, matching the performance of its classical NN counterpart. And the results indicate that, even in the presence of noise, this hybrid model can still achieve comparable performance to other ML methods.

Overall, the QNN outperforms the classical NN. To investigate whether this improvement is attributed to the quantum-enhanced feature embeddings, we calculated the correlation coefficient between the quantum-encoded features and PA values and compared it with that of the original features with PA values. To make a fair comparison, we chose the case of 32 features, 8 qubits/QC, 4 subencoders and 20 trainable parameters/QC. In this case, after the concatenation of the outputs of 4 subencoders, the number of the quantum-encoded features is still 32, identical to the number of the original features, ensuring the classical NN part in Figure 1b has the same neurons in each layer as the classical model which directly uses 32 original features. As shown in Figure 3, it is clear that the quantum-encoded features have strongly positive correlation with PA values and one quantum-encoded feature (feature index = 3) shows positive correlation which is 2 orders of magnitude stronger than other quantum-encoded features as well as these 32 original features. The results indicate that quantum circuits are capable of exactly capturing the feature representations in the complex high-dimensional space, thus being a good feature encoder to further improve classical NN models.

CONCLUSIONS

In this study, we present a predictive approach for PA using ML methods, integrating multiple descriptors of diverse types. The models demonstrated good prediction statistics via 5-fold cross validation. The combined Voting Regressor model outperforms individual models, achieving accuracy that approaches the experimental uncertainty of approximately ±2 kcal/mol. Next, we explored the capability of quantum circuits as feature encoders within neural network for PA regression tasks. In this hybrid model, we designed quantum circuits to encode classical features efficiently, reusing the circuits multiple times to maximize their encoding capacity while minimizing the required number of qubits. The results indicate that quantum circuits excel in capturing and processing high-dimensional feature space, thus facilitating the classical NN to achieve remarkable performances on PA predictions. Additionally, the results reveal that increasing the number of qubits, trainable gates, and quantum circuits generally enhances model performance, providing a clear pathway for optimization. This hybrid QNN model, despite using fewer trainable parameters, consistently outperforms its classical NN counterpart and several traditional ML methods. This highlights the significant potential of quantum circuits to improve ML performance, particularly in feature-rich applications. The study underscores the feasibility of leveraging quantum computing to complement classical ML, bridging the gap between quantum and classical paradigms. Currently we are still in the NISQ era and the limitations of QNN remain significant and cannot be ignored. For example, barren plateaus limit efficient pathway toward the optimal optimization of parameterized quantum circuits with deep depth. Meanwhile, the noise effects on hardware, the decoherence of physical qubits also prevent the availability of deep quantum circuits with large-scale qubits in QML. Considering these constraints, in this work we designed the quantum circuits with only 10 qubits. However, with the rapid advancements in quantum hardware and error mitigation techniques, deeper quantum circuits will be possible in the near future, and the advantages of quantum machine learning are expected to become more pronounced over time. And these developments on quantum computers will likely expand the applicability of hybrid models to a broader range of scientific and industrial problems, paving the way for innovative solutions in areas requiring high-dimensional data processing and predictive accuracy. All data and code are available at https://github.com/Neon8988/QNN_PA.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.4c01609.

The optimal hyperparameters of traditional models (PDF)

AUTHOR INFORMATION

Corresponding Author

Kenneth M. Merz, Jr − *Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, United States; Center for Computational Life Sciences, Lerner Research Institute, The Cleveland Clinic, Cleveland, Ohio 44106, United States;* orcid.org/0000-0001-9139-5893; Email: merz@chemistry.msu.edu

Author

Hongni Jin − *Department of Chemistry, Michigan State University, East Lansing, Michigan 48824, United States; Center for Computational Life Sciences, Lerner Research Institute, The Cleveland Clinic, Cleveland, Ohio 44106, United States;* orcid.org/0009-0009-4345-6300

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.4c01609

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support from the NIH (GM130641). The authors also thank the High Performance Computing Center (HPCC) at Michigan State University for providing all computational resources.

REFERENCES

(1) Berendsen, R. L.; Pieterse, C. M. J.; Bakker, P. A. H. M. The Rhizosphere Microbiome and Plant Health. *Trends Plant Sci.* **2012**, *17*, 478−486.

(2) Nicholson, J. K.; Connelly, J.; Lindon, J. C.; Holmes, E. Metabonomics: A Platform for Studying Drug Toxicity and Gene Function. *Nat. Rev. Drug Discovery* **2002**, *1*, 153−161.

(3) Griffin, J. L.; Bollard, M. E. Metabonomics: Its Potential as a Tool in Toxicology for Safety Assessment and Data Integration. *Curr. Drug Metab.* **2004**, *5*, 389−398.

(4) Mairinger, T.; Causon, T. J.; Hann, S. The Potential of Ion Mobility−Mass Spectrometry for Non-Targeted Metabolomics. *Curr. Opin. Chem. Biol.* **2018**, *42*, 9−15.

(5) Dodds, J. N.; Baker, E. S. Ion Mobility Spectrometry: Fundamental Concepts, Instrumentation, Applications, and the Road Ahead. *J. Am. Soc. Mass Spectrom.* **2019**, *30*, 2185−2195.

(6) Das, S.; Tanemura, K. A.; Dinpazhoh, L.; Keng, M.; Schumm, C.; Leahy, L.; Asef, C. K.; Rainey, M.; Edison, A. S.; Fernández, F. M.; Merz, K. M. *J. Am. Soc. Mass Spectrom.* **2022**, 33, 750−759.

(7) Aue, D. H.; Bowers, M. T.; Webb, H. M.; McIver, R. T. Equilibrium Constants for Gas-Phase Ionic Reactions. Accurate Determination of Relative Proton Affinities. *J. Am. Chem. Soc.* **1971**, *93*, 4314−4315.

(8) Nuttall, R. L.; Laufer, A. H.; Kilday, M. V. The Enthalpy of Formation of Ketene. *J. Chem. Thermodyn.* **1971**, *3*, 167−174.

(9) Chupka, W. A.; Berkowitz, J. Photoionization of Methane: Ionization Potential and Proton Affinity of CH4. *J. Chem. Phys.* **1971**, *54*, 4256−4259.

(10) Traeger, J. C. Heat of Formation for the Formyl Cation by Photoionization Mass Spectrometry. *Int. J. Mass Spectrom. Ion Processes* **1985**, *66*, 271−282.

(11) Berthelot, M.; Decouzon, M.; Gal, J. F.; Laurence, C.; Le Questel, J. Y.; Maria, P. C.; Tortajada, J. Gas-Phase Basicity and Site of Protonation of Polyfunctional Molecules of Biological Interest: FT-ICR Experiments and AM1 Calculations on Nicotines, Nicotinic Acid Derivatives, and Related Compounds. *J. Org. Chem.* **1991**, *56*, 4490−4494.

(12) French, M.; Kebarle, P. Pyrolysis of $C_2H_7^+$ and Other Ion−Molecule Reactions in Methane Containing Traces of Ethane. *Can. J. Chem.* **1975**, *53*, 2268−2274.

(13) Hunter, E. P. L.; Lias, S. G. Evaluated Gas Phase Basicities and Proton Affinities of Molecules: An Update. *J. Phys. Chem. Ref. Data* **1998**, *27b*, 413−656.

(14) Maksić, Z. B.; Kovačević, B.; Vianello, R. Advances in Determining the Absolute Proton Affinities of Neutral Organic Molecules in the Gas Phase and Their Interpretation: A Theoretical Account. *Chem. Rev.* **2012**, *112*, 5240−5270.

(15) Pople, J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. Gaussian−1 Theory: A General Procedure for Prediction of Molecular Energies. *J. Chem. Phys.* **1989**, *90*, 5622−5629.

(16) Curtiss, L. A.; Jones, C.; Trucks, G. W.; Raghavachari, K.; Pople, J. A. Gaussian−1 Theory of Molecular Energies for Second−row Compounds. *J. Chem. Phys.* **1990**, *93*, 2537−2545.

(17) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. Gaussian−2 Theory for Molecular Energies of First− and Second−row Compounds. *J. Chem. Phys.* **1991**, *94*, 7221−7230.

(18) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. Gaussian-3 (G3) Theory for Molecules Containing First and Second-Row Atoms. *J. Chem. Phys.* **1998**, *109*, 7764−7776.

(19) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 Theory. *J. Chem. Phys.* **2007**, *126* (8), 084108.

(20) Curtiss, L. A.; Raghavachari, K.; Pople, J. A. Gaussian−2 Theory Using Reduced Mo/Ller−Plesset Orders. *J. Chem. Phys.* **1993**, *98*, 1293−1298.

(21) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. Gaussian-3 Theory Using Reduced Mo/Ller-Plesset Order. *J. Chem. Phys.* **1999**, *110*, 4703−4709.

(22) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 Theory Using Reduced Order Perturbation Theory. *J. Chem. Phys.* **2007**, *127* (12), 124105.

(23) Barnes, E. C.; Petersson, G. A.; Montgomery, J. A.; Frisch, M. J.; Martin, J. M. L. Unrestricted Coupled Cluster and Brueckner Doubles Variations of W1 Theory. *J. Chem. Theory Comput.* **2009**, *5*, 2687−2693.

(24) Kolboe, S. Proton Affinity Calculations with High Level Methods. *J. Chem. Theory Comput.* **2014**, *10*, 3123−3128.

(25) Jin, H.; Merz, K. M. Modeling Zinc Complexes Using Neural Networks. *J. Chem. Inf. Model.* **2024**, *64*, 3140−3148.

(26) Jin, H.; Merz, K. M. Modeling Fe (II) Complexes Using Neural Networks. *J. Chem. Theory Comput.* **2024**, *20*, 2551−2558.

(27) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights into Chemical Systems. *Chem. Rev.* **2021**, *121*, 9816−9872.

(28) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3*, 434−443.

(29) Su, S.; Yang, Y.; Gan, H.; Zheng, S.; Gu, F.; Zhao, C.; Xu, J. Predicting the Feasibility of Copper(I)-Catalyzed Alkyne−Azide Cycloaddition Reactions Using a Recurrent Neural Network with a Self-Attention Mechanism. *J. Chem. Inf. Model.* **2020**, *60*, 1165−1174.

(30) Jin, H.; Merz, K. M. LigandDiff: De Novo Ligand Design for 3D Transition Metal Complexes with Diffusion Models. *J. Chem. Theory Comput.* **2024**, *20*, 4377−4384.

(31) Jin, H.; Merz, K. M., Jr Partial to Total Generation of 3D Transition-Metal Complexes. *J. Chem. Theory Comput.* **2024**, *20*, 8367−8377.

(32) Pang, C.; Qiao, J.; Zeng, X.; Zou, Q.; Wei, L. Deep Generative Models in *de Novo* Drug Molecule Generation. *J. Chem. Inf. Model.* **2024**, *64*, 2174−2194.

(33) Jin, H.; Merz, K. M. Toward AI/ML-Assisted Discovery of Transition Metal Complexes. *Annu. Rep. Comput. Chem.* **2024**, *20*, 225−267.

(34) Benedetti, M.; Lloyd, E.; Sack, S.; Fiorentini, M. Parameterized Quantum Circuits as Machine Learning Models. *Quantum Sci. Technol.* **2019**, *4*, 043001.

(35) Smaldone, A. M.; Batista, V. S. Quantum-to-Classical Neural Network Transfer Learning Applied to Drug Toxicity Prediction. *J. Chem. Theory Comput.* **2024**, *20*, 4901−4908.

(36) Bhatia, A. S.; Saggi, M. K.; Kais, S. Quantum Machine Learning Predicting ADME-Tox Properties in Drug Discovery. *J. Chem. Inf. Model.* **2023**, *63*, 6476−6486.

(37) Kao, P.-Y.; Yang, Y.-C.; Chiang, W.-Y.; Hsiao, J.-Y.; Cao, Y.; Aliper, A.; Ren, F.; Aspuru-Guzik, A.; Zhavoronkov, A.; Hsieh, M.-H.; Lin, Y.-C. Exploring the Advantages of Quantum Generative Adversarial Networks in Generative Chemistry. *J. Chem. Inf. Model.* **2023**, *63*, 3307−3318.

(38) Li, J.; Topaloglu, R. O.; Ghosh, S. Quantum Generative Models for Small Molecule Drug Discovery. *IEEE Trans. Quantum Eng.* **2021**, *2*, 1−8.

(39) Vitz, M.; Mohammadbagherpoor, H.; Sandeep, S.; Vlasic, A.; Padbury, R.; Pham, A. Hybrid Quantum Graph Neural Network for Molecular Property Prediction. *arXiv* **2024**.

(40) Ryu, J.-Y.; Elala, E.; Rhee, J.-K.-K. Quantum Graph Neural Network Models for Materials Search. *Materials* **2023**, *16*, 4300.

(41) Reddy, P.; Bhattacherjee, A. B. A Hybrid Quantum Regression Model for the Prediction of Molecular Atomization Energies. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 025019.

(42) Fourches, D.; Muratov, E.; Tropsha, A. But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189−1204.

(43) Landrum, G. *RDKit: open-source cheminformatics*, http://www.rdkit.org.

(44) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminf.* **2018**, *10* (1), 4.

(45) Besler, B. H.; Merz Jr, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990**, *11*, 431−439.

(46) Marenich, A. V.; Jerome, S. V.; Cramer, C. J.; Truhlar, D. G. Charge Model 5: An Extension of Hirshfeld Population Analysis for the Accurate Description of Molecular Interactions in Gaseous and Condensed Phases. *J. Chem. Theory Comput.* **2012**, *8*, 527−541.

(47) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H., et al., *Gaussian 16, Revision B.01*; Gaussian, Inc.: Wallingford CT, 2016.

(48) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.

(49) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminf.* **2015**, 7 (1), 20.

(50) Huang, H.-L.; Du, Y.; Gong, M.; Zhao, Y.; Wu, Y.; Wang, C.; Li, S.; Liang, F.; Lin, J.; Xu, Y.; et al. Experimental Quantum Generative Adversarial Networks for Image Generation. *Phys. Rev. Appl.* **2021**, *16*, 024051.

(51) Anagolum, S.; Alavisamani, N.; Das, P.; Qureshi, M.; Shi, Y. Elivagar: Efficient Quantum Circuit Search for Classification. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*; ACM: New York, NY, USA, 2024; pp. 336−353.

(52) Wang, H.; Ding, Y.; Gu, J.; Lin, Y.; Pan, D. Z.; Chong, F. T.; Han, S. Quantumnas: Noise-adaptive search for robust quantum circuits. In *2022 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*; IEEE, 2022, pp. 692−708.

(53) Javadi-Abhari, A.; Treinish, M.; Krsulich, K.; Wood, C. J.; Lishman, J.; Gacon, J.; Martiel, S.; Nation, P. D.; Bishop, L. S.; Cross, A. W.et al. Quantum computing with Qiskit, *arXiv*, **2024**. .

(54) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.