Quantum QSAR for drug discovery

Alejandro Giraldo [⋆1], Daniel Ruiz, Mariano Caruso [2,3,4], and Guido Bellomo [5]

[1] QNOW Technologies, Delaware, USA
alejandro@qnow.tech, daniel@qnow.tech
[2] UGR, Granada, Spain
[3] UNIR, La Rioja, Spain
[4] FIDESOL, Granada, Spain
mcaruso@fidesol.org
[5] CONICET - UBA, ICC, Argentina
gbellomo@icc.fcen.uba.ar

Abstract. Quantitative Structure-Activity Relationship (QSAR) modeling is key in drug discovery, but classical methods face limitations when handling high-dimensional data and capturing complex molecular interactions. This research proposes enhancing QSAR techniques through Quantum Support Vector Machines (QSVMs), which leverage quantum computing principles to process information in Hilbert spaces. By using quantum data encoding and quantum kernel functions, we aim to develop more accurate and efficient predictive models.

Keywords: QSAR, classification, drug discovery, support vector machines, quantum kernel.

1 Introduction

QSAR models aim to establish relationships between the physicochemical properties of compounds and their molecular structures. Hansch and Fujita (1964) These mathematical models serve as valuable tools in pharmacological studies by providing an *in silico* methodology to test and classify new compounds with desired properties, diminish the need for laboratory experimentation Natarajan, Natarajan, and Basak (2025). QSAR models are used, for example, to predict pharmacokinetic processes such as absorption, distribution, metabolism, and excretion, ADME, which refers to the processes that describe how a drug or chemical substance moves through and is processed by the body.

In other fields, (QSAR) plays an important role; for example, *in silico* toxicity studies have become fundamental in drug development. A prevalent way in which QSAR is used is in this context of prediction, which helps us understand how we can link toxicity outcomes to the structural properties of specific compounds.

Many models show decent performance throughout their implementations, as they rely on a pipeline that is optimizable and improvable, whereas machine

⋆ corresponding author: alejandro@qnow.tech

learning methods will always involve a tradeoff between accuracy and interpretability.

1.1 Evolution of QSAR Modeling Approaches

Traditionally, QSAR relied on linear regression models, but these were quickly replaced by more sophisticated approaches. Bayesian neural networks emerged as a powerful alternative, demonstrating the ability to distinguish between drug-like and non-drug-like molecules with high accuracy Ajay, Walters, and Murcko (1998). These models showed excellent generalization capabilities, correctly classifying more than 90% of the compounds in the database while maintaining low false positive rates.

Random forest algorithms have also proven to be effective tools for QSAR modeling Svetnik et al. (2003). This ensemble method, which combines multiple decision trees, has shown superior performance in predicting biological activity based on molecular structure descriptors. Its advantages include built-in performance evaluation, descriptor importance measures, and compound similarity computations weighted by the relative importance of descriptors.

In general, the process involves three main stages: obtaining a training dataset with measured properties of known compounds, encoding information about the compounds' structure, and building a model to predict properties from the encoded structural data, followed by training the model. **(1.)** Preprocessing and extraction of molecular descriptors. **(2.)** Encoding of classical data into quantum states using a feature map. **(3.)** Classification using support vector machines (SVM) with classical and quantum kernels.

1.2 General Pipeline

1. Compound Collection and Curation: The process begins with the collection of candidate compounds, either from experimental or theoretical sources. These compounds are curated to ensure suitability for the selected biological target. This step may involve filtering based on physicochemical properties or prior biological knowledge.

2. Data Preprocessing and Descriptor Calculation: Regardless of the target, all data undergoes preprocessing to normalize and standardize values. Molecular descriptors (features) are computed for each compound. These may include physicochemical properties (e.g., molecular weight, hydrogen bond donors /acceptors, rotatable bonds) and structural fingerprints (e.g., MACCs, ECFP). Given the constraints of current quantum hardware, dimensionality reduction is often necessary. Techniques such as Principal Component Analysis (PCA) are applied to retain the most informative components while reducing the number of features, thus minimizing the required number of qubits for quantum encoding.

3. Dataset Balancing and Partitioning: In this study, the dataset is inherently imbalanced and relatively small. Although advanced balancing techniques (e.g., SMOTE, RandomUndersampling) were not applied, the dataset serves as a practical testbed for rapid experimentation and for evaluating the

methodology across different data volumes. For future work, balancing strategies could be incorporated to assess their impact on model performance.

4. Classical-to-Quantum Data Mapping: Once the dataset is enriched and preprocessed, classical features are mapped to quantum states using a feature map (e.g., ZZFeatureMap). The number of qubits required is determined by the dimensionality of the reduced feature set. This mapping is a critical step, as it enables the exploitation of quantum space. Schuld and Killoran (2018)

5. Model Training and Evaluation: The enriched dataset is used to train both classical and quantum models. For quantum models, the support vector machine (SVM) leverages quantum kernels Li et al. (2019), with training and inference performed either on quantum simulators or real quantum processing units (QPUs). The choice of platform and the number of qubits used are dictated by the final feature dimensionality and hardware availability.

Experiments are typically partitioned into training and test sets, with performance metrics (e.g., accuracy) computed to compare classical and quantum approaches.

6. Scalability and Implementation Notes: Current quantum hardware imposes strict limits on the number of qubits and circuit depth, constraining the size and complexity of datasets that can be processed. Execution time and noise are also significant factors, especially when running on real QPUs. These limitations highlight the importance of dimensionality reduction and motivate ongoing research into error mitigation and hybrid quantum-classical workflows.

While this work focuses on quantum SVMs, alternative quantum approaches such as Variational Quantum Circuits (VQCs) could be explored in future studies to further assess the potential of quantum machine learning in QSAR applications.

This detailed pipeline description aims to clarify the methodological steps, justify key design choices, and provide a foundation for reproducibility and future scalability assessments.

1.3 Dataset, Descriptors, and Features

Each candidate molecule has a series of molecular descriptors or *features*, such as the median effective concentration, Lipinski descriptors, which are a type of molecular fingerprint—specifically a 2D structure fingerprint of 166 bits—used to represent and compare molecular structures.

These descriptors may include records of experimental results; for example, the concentration $\texttt{EC}_{50}$ indicates the amount of a compound required to elicit 50% of the maximum biological effect after a specific exposure time. This is expressed in molar units M: `mol/L`.

Each dataset requires data processing. In models involving `ADME`, it is important to work with concentration parameters that provide meaningful information. Concentrations are usually in the nM range, and a logarithmic transformation is employed, defining the potential of this concentration as $p\texttt{EC}_{50}$, denoted by $p = -\log_{10}\left(\texttt{EC}_{50} \times 10^{-9}\right)$, which facilitates its use in quantitative analyses of biological activity Neubig (2003).

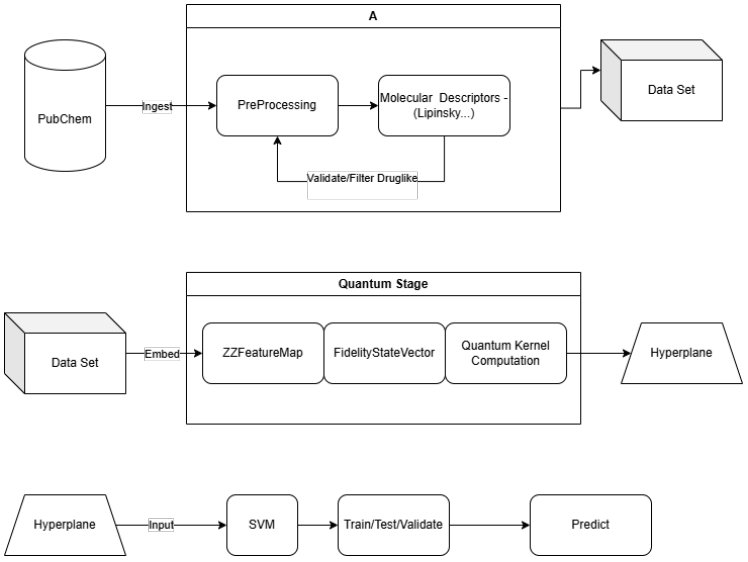

Fig. 1. High level pipeline from data perspective

To contextualize these models within a study domain, we will use a dataset where the target is the `M2 Muscarinic Acetylcholine` receptor, a G protein-coupled receptor that plays a crucial role in the parasympathetic nervous system, particularly in regulating cardiac function and smooth muscle activity. It is encoded by the `CHRM2` gene in humans.

In the pharmacokinetic context, we will use Lipinski's *rule of five*, which is a set of empirical criteria fundamental to drug design. It describes molecular properties relevant to pharmacokinetics in the human body, including absorption, distribution, metabolism, and excretion (`ADME`). This rule helps assess the likelihood that a chemical compound exhibits adequate pharmacokinetic properties for oral administration in humans, based on four key molecular properties: molecular weight ($\leq$ 500`Da`), number of hydrogen bond donors ($\leq$ 5), number of hydrogen bond acceptors ($\leq$ 10), and octanol-water partition coefficient ($\log P \leq 5$). A compound that meets at least three of these criteria is more likely to have good oral bioavailability.

It is important to note that, while this rule is useful for predicting pharmacokinetic properties, it does not predict whether a compound will be pharmacologically active. Its main utility lies in the early stages of drug discovery, allowing researchers to filter out compounds with a low probability of success before conducting costly experiments.

From the structural information of the molecules, various descriptors of interest are extracted, among which the following are highlighted: number of hydrogen bond donors, n_d, representing the number of functional groups that can donate a hydrogen atom; number of hydrogen bond acceptors, n_a, which counts the num-

ber of sites capable of accepting a hydrogen atom; number of rotatable bonds, ρ, indicating molecular flexibility associated with the ability to rotate around single bonds; molecular weight, w, which defines the mass of the molecule in atomic mass units.

In this initial study, we consider a limited number of descriptors, which will depend on the experiments described later, associated with the representational power of classical data in quantum systems. For practical purposes, these features were defined as part of a potentially more refined feature engineering process.

1.4 Data Processing for the Model

Regarding data processing, we aim to maintain a consistent scale of values to enable operations, standardize values, and in some cases compress data. Therefore, proper preparations will be carried out before training any model. In this way, the feature vector is given by $\boldsymbol{x} = (n_d, n_a, \rho, w, \cdots)$. Due to the variability in numerical scales of these descriptors, normalization is performed using the `minmax` method, so that each component l of the rescaled vector, $\boldsymbol{x}'$, is expressed as $x'_l = \big(x_l - \texttt{min}\{x_l\}\big)/\big(\texttt{max}\{x_l\} - \texttt{min}\{x_l\}\big)$, e.g., $\forall l$: $x'_l \in [0, 1]$.

2 Classical and Quantum Models

In the context of supervised machine learning, we work with labeled data, particularly a training dataset of size N, $\{(\boldsymbol{x}_i, y_i)\}_{i \in I_N}$, where $\boldsymbol{x}_i$ is the feature vector and y_i its corresponding label indicating whether it is suitable or not. The goal is to find a predictor for y from a family of parameterized predictors with a real-valued parameter vector $\boldsymbol{q}$, by solving an optimization problem for a function of $\boldsymbol{q}$. Specifically, we consider regression and classification models, with predictor families defined respectively by the following functions f and g:

$$f(\boldsymbol{x}, \boldsymbol{q}) = \sum_{k=1}^{m} q_k \phi_k(\boldsymbol{x}), \quad g(\boldsymbol{x}, \boldsymbol{\alpha}, b) = \texttt{sgn}\left[\sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right]$$

where in the second case the parameter vector is $\boldsymbol{q} = (\boldsymbol{\alpha}, b)$. We aim to estimate whether a given compound is suitable using $\hat{y}$, which corresponds to the output of the respective trained predictors. The functions $\phi_k(\boldsymbol{x})$ are called *feature maps*, and they are used to transform the features $\boldsymbol{x}$ to another space—either of lower dimensionality or one that reveals separability between two given points $\boldsymbol{x}, \boldsymbol{x}'$. The function $K(\boldsymbol{x}, \boldsymbol{x}')$ is called a *kernel*, and its dependence on $(\boldsymbol{x}, \boldsymbol{x}')$ arises through the feature maps $(\phi(\boldsymbol{x}), \phi(\boldsymbol{x}'))$ Huang et al. (2022). Some examples include the linear kernel $\phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{x}')$, the polynomial kernel $(\phi(\boldsymbol{x}) \cdot \phi(\boldsymbol{x}') + c)^d$, or the Gaussian kernel $\exp[-\gamma||\phi(\boldsymbol{x}) - \phi(\boldsymbol{x}')||^2]$.

Since the data is classical, quantum computing could add value in two parts of the process: **1.** solving the optimization problems underlying the training phase, and **2.** encoding data using quantum kernels. The data is encoded into quantum states through a feature map implemented by a unitary operator $U(\boldsymbol{x})$,

giving the representation $|\phi(\boldsymbol{x})\rangle = U(\boldsymbol{x})\,|0\rangle^{\otimes n}$. The similarity between two quantum states is measured using fidelity, and the quantum kernel is defined as $K_q(\boldsymbol{x}, \boldsymbol{x'}) = |\langle\phi(\boldsymbol{x})|\phi(\boldsymbol{x'})\rangle|^2$. This kernel naturally incorporates superposition and entanglement effects, enabling the capture of complex nonlinear relationships in feature space Huang et al. (2021). Classification is then carried out by training an `SVM` where the classical kernel is replaced by the quantum kernel K_q. This approach allows us to explore the efficiency and potential of quantum algorithms in `QSAR` scenarios, comparing them with classical approaches Havlíček et al. (2019). The underlying optimization problem in both regression and classification models involves a quadratic problem that can be solved using classical algorithms like gradient descent, heuristic methods like simulated annealing and its quantum variant, or by using gate-based quantum computing to implement algorithms such as `VQE` or `QAOA`.

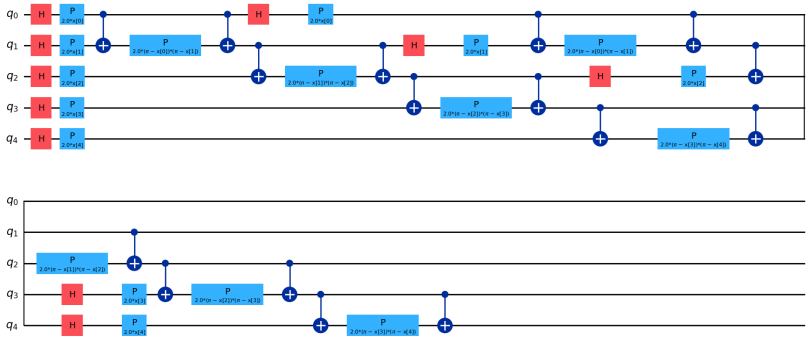

Fig. 2. ZZFeaturemap as a linear entangled quantum kernel.

3 Results and Discussion

In this section, we present the results obtained from implementing the regression and classification models. These models can be deployed on either classical or quantum hardware. In particular, for the quantum setting, this includes annealing-based computers such as those from `DWave`, or gate-based universal quantum computers like those developed by `IBM`.

To compare the performance across different models, we have chosen the metric known as accuracy, defined as:

$$\texttt{acc} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[\hat{y}_i = y_i], \tag{1}$$

where n is the number of test samples, and $\mathbf{1}[\hat{y}_i = y_i]$ is the indicator function over the set of correct predictions.

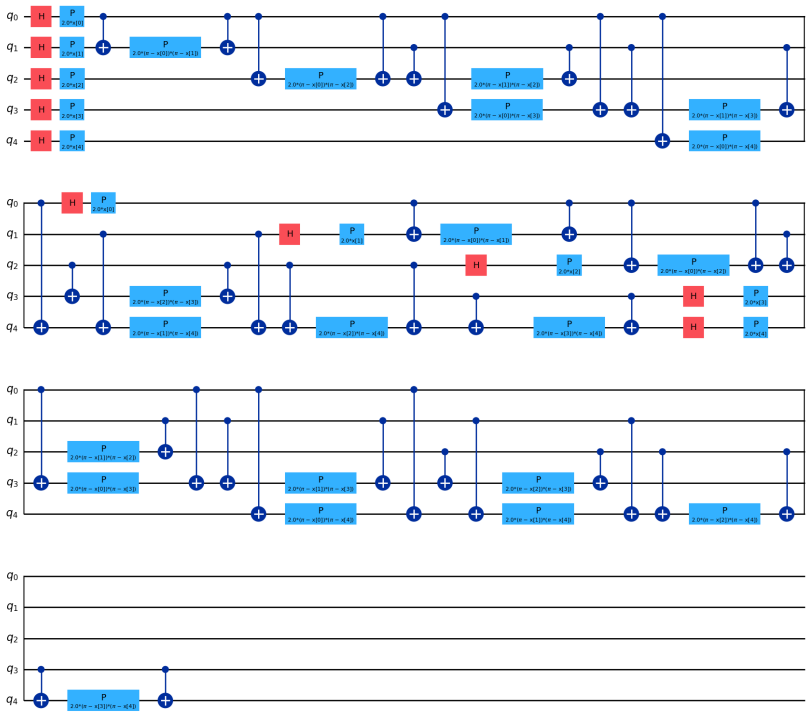

Fig. 3. ZZFeaturemap as a full entangled quantum kernel.

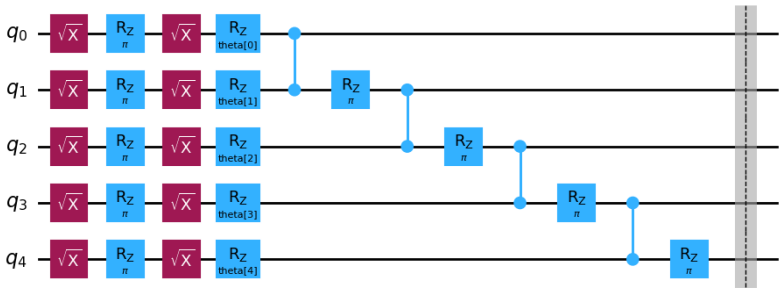

Fig. 4. Custom linear entangled quantum kernel.

A comparison of the different regression (REG) and classification (SVM) models, based on whether classical or quantum algorithms were used, is summarized in Table 1.

Table 1. We denote by ***c*** and ***q*** the classical and quantum terms, respectively, to qualify the type of model or the kernel as appropriate. The acronyms sim and QPU refer to execution on quantum simulators and real quantum processors, respectively.

model	type	acc	execution	kernel
$\texttt{REG}_1$	c	0.95	CPU	–
$\texttt{REG}_2$	q	0.97	sim	–
$\texttt{SVM}_1$	c	0.87	CPU	c | linear
$\texttt{SVM}_2$	c/q	0.98	sim	q | linear - Fig 2
$\texttt{SVM}_3$	c/q	0.83	sim	q | nolinear - Fig 3
$\texttt{SVM}_4$	c/q	0.40	QPU	q |linear Fig 4

4 Conclusions

A pipeline has been developed that integrates traditional QSAR methods with quantum machine learning techniques. The methodology includes preprocessing and normalization of molecular descriptors, projection of these data into quantum states via the ZZ-feature map, and classification using SVM with both classical and quantum kernels. This approach allows for the evaluation of the potential of quantum methods to improve classification in chemico-pharmaceutical applications, relying on rigorous mathematical foundations and the emerging capabilities of quantum computing.

The potential advantages of this integration lie in the ability of quantum kernels to capture complex correlations, even in scenarios with limited data, which may translate into improvements in performance compared to classical techniques.

Acknowledgment

This work was supported by the project ECO – 20241014 QUORUM funded by Ministerio de Ciencia, Innovación y Universidades, through CDTI.

References

Ajay, Walters, W. P., & Murcko, M. A. (1998). Can we learn to distinguish between drug-like and nondrug-like molecules? *Journal of Medicinal Chemistry*, *41*(18), 3314–3324. doi: 10.1021/jm970666c

Hansch, C., & Fujita, T. (1964). ρ-σ-π analysis. a method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, *86*(8), 1616–1626. Retrieved from https://doi.org/10.1021/ja01062a035 doi: 10.1021/ja01062a035

Havlíček, V., Córcoles, A. D., Temme, K., Harrow, A. W., Kandala, A., Chow, J. M., & Gambetta, J. M. (2019). Supervised learning with quantum-enhanced feature spaces. *Nature*, *567*(7747), 209–212. Retrieved from https://doi.org/10.1038/s41586-019-0980-2 doi: 10.1038/s41586-019-0980-2

Huang, H.-Y., Broughton, M., Cotler, J., Chen, S., Li, J., Mohseni, M., ... McClean, J. R. (2022). Quantum advantage in learning from experiments. *Science*, *376*(6598), 1182–1186. Retrieved from https://www.science.org/doi/10.1126/science.abn7293 doi: 10.1126/science.abn7293

Huang, H.-Y., Broughton, M., Mohseni, M., Babbush, R., Boixo, S., Neven, H., & McClean, J. R. (2021). Power of data in quantum machine learning. *Nature Communications*, *12*(1), 2508. Retrieved from https://doi.org/10.1038/s41467-021-22539-9 doi: 10.1038/s41467-021-22539-9

Li, K., et al. (2019). Quantum-inspired support vector machine. *arXiv preprint arXiv:1906.08902*. (arXiv:1906.08902 [cs.LG])

Natarajan, R., Natarajan, G. S., & Basak, S. C. (2025). Quantitative structure–activity relationship (QSAR) modeling of chiral CCR2 antagonists with a multidimensional space of novel chirality descriptors. *Molecules*, *30*(2), 307. Retrieved from https://doi.org/10.3390/molecules30020307 doi: 10.3390/molecules30020307

Neubig, R. R. (2003). International union of pharmacology committee on receptor nomenclature and drug classification. xxxviii. update on terms and symbols in quantitative pharmacology. *Pharmacological Reviews*, *55*(4), 597–606. Retrieved from https://doi.org/10.1124/pr.55.4.4 doi: 10.1124/pr.55.4.4

Schuld, M., & Killoran, N. (2018). Quantum machine learning in feature hilbert spaces. *arXiv preprint arXiv:1803.07128*. (arXiv:1803.07128 [quant-ph])

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: A classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1947–1958.