

Q²SAR: overcoming classical bottlenecks in drug discovery via quantum multiple kernel learning

Alejandro Giraldo¹ , Daniel Ruiz¹ , Mariano Caruso^{2,3,4} , and Guido Bellomo⁵ 

¹ QNOW Technologies, Delaware, USA

{alejandro,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 a foundational computational methodology in early-stage drug discovery, heavily relied upon for predicting compound toxicity, bioavailability, and therapeutic potential. However, classical methods often struggle to effectively map the highly complex, non-linear, and high-dimensional interactions inherent in molecular data, leading to reduced predictive accuracy and costly late-stage clinical failures. In this paper, we present a Quantum Multiple Kernel Learning (QMKL) framework—dubbed Next-Gen Q²SAR—that leverages Quantum Support Vector Machines (QSVMs) to overcome these classical limitations. By encoding molecular descriptors into exponentially large quantum Hilbert spaces, our approach substantially enhances the expressiveness of non-linear modeling. Benchmarking our quantum-enhanced framework on a dataset targeting the DYRK1A kinase (a critical target for Alzheimer’s disease), the QMKL-SVM achieves an impressive Area Under the Curve (AUC) score of 0.8750, significantly outperforming classical state-of-the-art Gradient Boosting models (AUC = 0.8037). Furthermore, we establish a theoretical and empirical pathway toward resolving classical data bottlenecks through projected quantum kernels (PQK) and measurement accelerators. As quantum computing architecture matures, this framework paves the way for autonomous cognitive architectures and self-improving drug discovery pipelines, promising to unlock deeper insights across vast chemical spaces and to accelerate the development of life-saving therapeutics.

Keywords: qsar , drug discovery , quantum multiple kernel learning , support vector machines , dyrk1a

Q²SAR: superando cuellos de botella clásicos en el descubrimiento de fármacos vía quantum multiple kernel learning

Abstract. El modelado de Relaciones Cuantitativas Estructura-Actividad (QSAR) es una metodología computacional fundamental en las fases tempranas del descubrimiento de fármacos, de la cual se depende enormemente para predecir la toxicidad, biodisponibilidad y potencial terapéutico de los compuestos. Sin embargo, los métodos clásicos a menudo tienen dificultades para mapear eficazmente las interacciones altamente complejas, no lineales y de alta dimensionalidad inherentes a los datos moleculares, lo que conduce a una precisión predictiva reducida y a costosos fracasos clínicos en etapas tardías. En este artículo, presentamos un marco de trabajo de *Quantum Multiple Kernel Learning* (QMKL) —denominado Q²SAR de Próxima Generación— que aprovecha las Máquinas de Vectores de Soporte Cuánticas (QSVMs) para superar estas limitaciones clásicas. Al codificar los descriptores moleculares en espacios de Hilbert cuánticos exponencialmente grandes, nuestro enfoque mejora sustancialmente la expresividad del modelado no lineal. Evaluando nuestro marco mejorado cuánticamente en un conjunto de datos dirigido a la quinasa DYRK1A (una diana terapéutica crítica para la enfermedad de Alzheimer), el QMKL-SVM logra una impresionante puntuación de Área Bajo la Curva (AUC) de 0.8750, superando significativamente a los modelos clásicos del estado del arte basados en *Gradient Boosting* (AUC = 0.8037). Además, establecemos una vía teórica y empírica para resolver los cuellos de botella de datos clásicos mediante kernels cuánticos proyectados (*Projected Quantum Kernels*, PQK) y aceleradores de medición. A medida que la arquitectura de computación cuántica madura, este marco de trabajo allana el camino para arquitecturas cognitivas autónomas y canales de descubrimiento de fármacos con mejora automática, prometiendo desbloquear conocimientos más profundos en vastos espacios químicos y acelerar el desarrollo de terapias que salvan vidas.

Keywords: qsar , descubrimiento de fármacos , quantum multiple kernel learning , dyrk1a

1 Introducción

La predicción precisa de la actividad biológica para compuestos químicos novedosos y no caracterizados es un desafío definitorio en la química computacional y la farmacología modernas. Esta tarea fundamental forma el núcleo del modelado de Relaciones Cuantitativas Estructura-Actividad (QSAR), una metodología crítica que busca inferir los efectos biológicos y toxicológicos de las moléculas

mediante el análisis de patrones intrincados entre sus descriptores moleculares y resultados experimentales conocidos (Burbidge et al., 2001). Los modelos QSAR tradicionales juegan un papel esencial en la predicción de la biodisponibilidad, los efectos fuera de la diana (*off-target*) y el potencial terapéutico general de los fármacos candidatos, intentando, en última instancia, reducir los costos astronómicos y los largos plazos asociados con las altas tasas de deserción en las últimas etapas del desarrollo de fármacos.

A pesar de su adopción industrial generalizada, los enfoques QSAR clásicos enfrentan profundas limitaciones. Los datos químicos y biológicos son intrínsecamente de alta dimensionalidad y altamente no lineales. Los algoritmos clásicos típicamente dependen de técnicas agresivas de reducción de dimensionalidad (como el Análisis de Componentes Principales o PCA) o de representaciones heurísticas poco profundas que, aunque tratables matemáticamente, a menudo descartan información relacional estructural y semántica crucial. En consecuencia, los modelos basados en paradigmas clásicos como *Random Forests* o Perceptrones Multicapa estándar a menudo se encuentran con un techo de rendimiento intrínseco, sufriendo de una generalizabilidad reducida y una precisión predictiva mermada cuando se enfrentan a conjuntos de datos pequeños o altamente desbalanceados.

Para superar estas barreras, introducimos un marco de modelado mejorado cuánticamente que integra Máquinas de Vectores de Soporte Cuánticas (QSVMs) y *Quantum Multiple Kernel Learning* (QMKL). El *Quantum Machine Learning* (QML) propone un cambio del uso de datos o recursos computacionales de naturaleza cuántica. Dado que los datos de nuestro problema base son *clásicos*, es necesario mapearlos a los espacios de Hilbert de los sistemas cuánticos. A través de *feature maps* cuánticos especializados —representados por una transformación unitaria $U(x)$ — los descriptores moleculares clásicos x se codifican en un estado cuántico $|\phi(x)\rangle = U(x)|0^n\rangle$. El procesamiento de datos en estos espacios representacionales de alta dimensionalidad permite el descubrimiento de patrones estructura-actividad intrincados y no lineales que son inaccesibles para los algoritmos clásicos.

Nuestra metodología propuesta es agnóstica a dianas moleculares específicas y se generaliza a través de múltiples dominios terapéuticos. Como una demostración tangible de rendimiento mejorado y ventaja cuántica práctica, aplicamos nuestro enfoque Q²SAR a la identificación de inhibidores para la quinasa DYRK1A, una diana terapéutica de gran relevancia en el tratamiento de trastornos neurodegenerativos, particularmente la enfermedad de Alzheimer. Al evaluar el rendimiento de combinaciones de kernels clásicos y cuánticos ponderados óptimamente, nuestro enfoque híbrido demuestra un salto significativo en la capacidad predictiva, evidenciado por métricas superiores de la característica operativa del receptor (ROC) en comparación con los modelos clásicos de referencia.

Además, abordamos los desafíos apremiantes de la era *Noisy Intermediate-Scale Quantum* (NISQ), específicamente el "cuello de botella de acceso a datos" y la sobrecarga computacional de la ejecución repetitiva de circuitos. Al integrar aceleraciones algorítmicas novedosas —como *Projected Quantum Kernels* (PQK)

y combinaciones lineales de unitarios (*Linear Combination of Unitaries*, LCU) para la estimación del valor esperado en un solo *shot*— presentamos un marco rigurosamente justificado que va más allá de las promesas asintóticas heurísticas hacia una aplicación industrial demostrablemente viable.

2 Metodología

Para afirmar de manera convincente que una propuesta cuántica representa una mejora sobre el estado del arte, es imperativo ir más allá de las justificaciones basadas únicamente en la complejidad de consultas y adoptar un marco de trabajo de extremo a extremo (*end-to-end*). Nuestra metodología une los flujos de trabajo de la quimioinformática clásica con los métodos de *kernel* cuánticos.

2.1 Evolución de los modelos de clasificación: SVM $\rightarrow$ QSVM

La Máquina de Vectores de Soporte clásica (SVM) ha sido, durante décadas, un pilar fundamental en las tareas de clasificación debido a su solidez teórica basada en la maximización del margen de separación entre clases. Dado un conjunto de datos de entrenamiento compuesto por M muestras $\{(x_i, y_i)\}_{i=1, \dots, M}$, donde $x_i \in \mathbb{R}^N$ son los vectores de características y $y_i \in \{-1, 1\}$ son las etiquetas, el problema de optimización primal busca encontrar el hiperplano óptimo caracterizado por un vector normal w y un sesgo b :

$$\min_{w, b} \frac{1}{2} \|w\|^2 \quad (1)$$

sujeto a las restricciones $y_i(w \cdot x_i + b) \geq 1 \quad \forall i \in 1, \dots, M$.

Para manejar datos que no son linealmente separables, se emplea el *kernel trick*, mapeando los datos en un espacio de mayor dimensionalidad a través de una función $\phi(x)$. La función del *kernel* se define como el producto interno:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad (2)$$

A medida que escalan la dimensionalidad de los datos N y el número de ejemplos M , evaluar la matriz del *kernel* clásicamente se vuelve computacionalmente prohibitivo. La Máquina de Vectores de Soporte Cuántica (QSVM) surge para explotar el paralelismo cuántico utilizando un circuito cuántico parametrizado para implementar el *Feature Map* $\phi(x)$, procesando la información directamente en el espacio de Hilbert cuántico.

2.2 Flujo de trabajo para Q²SAR

Los modelos QSAR clásicos a menudo fallan al intentar capturar interacciones altamente no lineales y correlaciones ocultas en conjuntos de datos farmacológicos con alta dimensionalidad y clases desbalanceadas. Para superar este techo de rendimiento, proponemos un riguroso flujo de trabajo denominado Quantum QSAR (Q²SAR):

1. **Curación y transformación:** Conversión de cadenas **SMILES** en características numéricas integrales (descriptores fisicoquímicos, geométricos y electrónicos).
2. **Reducción de dimensionalidad:** Aplicación de Análisis de Componentes Principales (PCA) para retener la máxima varianza, adaptando explícitamente los datos al número de qubits disponibles en la Unidad de Procesamiento Cuántico (QPU).
3. **Mapeo cuántico:** Proyección de los vectores reducidos en el espacio de Hilbert mediante circuitos entrelazados.
4. **Clasificación SVM:** Cálculo de la matriz de *kernel* combinada y optimización.

3 Arquitectura cuántica

3.1 Codificación de datos cuánticos: ZZFeatureMap

La eficacia de una **QSVM** depende críticamente de la estrategia de mapeo de características. En los dispositivos de la era **NISQ**, la codificación de amplitud (*Amplitude Encoding*) se descarta porque requiere circuitos exponencialmente profundos (a través de **QRAM**), introduciendo un ruido insalvable. En su lugar, nuestra arquitectura adopta estrategias de codificación por ángulo (*Angle Encoding*), específicamente el **ZZFeatureMap**.

Esta técnica transforma el vector de datos clásicos x en un estado cuántico $|\psi(x)\rangle$ aplicando una capa de compuertas de Hadamard, seguida de rotaciones locales en el eje Z parametrizadas por las características individuales (por ejemplo, $R_z(x_i)$) y, lo más importante, compuertas de entrelazamiento **ZZ**. La interacción **ZZ** es crucial porque está parametrizada por productos cruzados de las características (típicamente $x_i \cdot x_j$). Esta operación matemática proyecta directamente los datos en un espacio de Hilbert donde las correlaciones de orden superior (interacciones no lineales entre descriptores moleculares) se capturan de forma nativa. Mientras que los modelos clásicos luchan por modelar estas combinaciones no lineales explícitamente sin explotar dimensionalmente, el **ZZFeatureMap** codifica esta topología compleja eficientemente, dotando al *Ansatz* de una ventaja de expresividad medible.

El *kernel* de fidelidad cuántica (FQK) mide entonces el solapamiento global entre dos estados parametrizados:

$$K_{FQK}(x_i, x_j) = |\langle \psi(x_i) | \psi(x_j) \rangle|^2 \quad (3)$$

3.2 Mitigación de barren plateaus: projected quantum kernels

A medida que aumenta el número de qubits, el espacio de Hilbert se vuelve tan vasto que estados arbitrarios tienden a ser ortogonales. Esto produce una matriz de *kernel* que se concentra exponencialmente hacia la matriz identidad, anulando drásticamente la capacidad de aprendizaje del modelo (un fenómeno

estrechamente relacionado con los *barren plateaus* en el entrenamiento de kernels).

Para combatir esto, utilizamos Kernels Cuánticos Proyectados (PQK). En lugar de medir la fidelidad global del estado, el PQK extrae valores esperados locales proyectando sobre matrices de densidad reducidas (1-RDM o 2-RDM):

$$K_{PQK}(x_i, x_j) = \exp \left[-\gamma \sum_k \sum_{P \in \{X, Y, Z\}} (\text{tr}[P\rho_k(x_i)] - \text{tr}[P\rho_k(x_j)])^2 \right]. \quad (4)$$

Esta proyección reduce drásticamente la sensibilidad al ruido del hardware, evita la concentración exponencial y mejora notablemente la generalización del modelo.

3.3 Quantum Multiple Kernel Learning (QMKL)

En lugar de depender de un único *kernel* cuántico (que podría ser sensible a fluctuaciones de ruido o sesgos del hardware), el QMKL aprende una combinación lineal óptima de un conjunto predefinido de *kernels* base $\{K_i\}_{i=1, \dots, m}$:

$$K = \sum_{i=1}^m w_i K_i, \quad \text{con } w_i \geq 0, \quad \sum_{i=1}^m w_i = 1 \quad (5)$$

Al optimizar algorítmicamente los pesos w_i para una mezcla de matrices clásicas robustas (como *Random Forest Kernel* o RBF) y *kernels* cuánticos expresivos, explotamos las características cuánticas solo donde aportan valor, respaldados por la solidez estadística clásica donde el espacio es linealmente manejable.

3.4 Aceleraciones algorítmicas en NISQ: HoLCUs y Pegasos

Evaluar miles de circuitos repetitivos para construir la matriz de Gram supone un cuello de botella logístico en la era NISQ. Para mitigar esto, empleamos el algoritmo HoLCUs (*Hamiltonian evaluation of Linear Combination of Unitaries*). HoLCUs permite medir el valor esperado de cualquier operador no unitario utilizando un solo circuito cuántico y un qubit auxiliar (*ancilla*). Al proyectar las 1-RDMs en una sola pasada, HoLCUs evita enviar miles de circuitos individuales a la cola de la QPU, logrando una aceleración temporal (*speedup* en tiempo real) de hasta $22.5\times$. Además, utilizamos el algoritmo Pegasos – QSVM, que optimiza el SVM en su formulación primal mediante descenso de gradiente estocástico, reduciendo dramáticamente las exigencias de memoria y mitigando el ruido de *shot* al muestrear subconjuntos aleatorios.

4 Resultados

La evaluación empírica de nuestra metodología Q²SAR examina su viabilidad más allá de los argumentos heurísticos. Para garantizar una evaluación justa,

el modelo cuántico se comparó directamente con métodos clásicos de última generación entrenados sobre los mismos descriptores tras el procesamiento PCA.

4.1 Benchmark para DYRK1A

La identificación de inhibidores para la quinasa DYRK1A (una diana terapéutica crítica en el alzhéimer) es un caso de uso con inmenso impacto social. Este conjunto de datos farmacológico presentaba desafíos clásicos: clases desbalanceadas (muy pocos inhibidores activos en comparación con moléculas inactivas) y una alta complejidad estructural en sus descriptores.

Los métodos clásicos del estado del arte, como *Random Forest* (RF) y *Gradient Boosting* (GB), sufren frente a este desbalance debido a su tendencia a crear árboles de decisión sesgados hacia la clase mayoritaria. En la clasificación de *Gradient Boosting*, se alcanzó un AUC de 0.8037. Aunque competitivo, este modelo clásico evidenció un límite natural al intentar segmentar las interacciones de orden superior.

En contraste, el modelo QMKL-SVM logró una asombrosa puntuación AUC de 0.8750. Este salto de casi 7 puntos porcentuales no es marginal; representa una transformación en la capacidad diagnóstica. Al utilizar la métrica AUC (Área Bajo la Curva ROC), estamos evaluando la robustez del modelo independientemente del umbral de decisión, demostrando que el modelo cuántico clasifica sistemáticamente las moléculas verdaderamente activas por encima de las inactivas con mayor fiabilidad.

Quizás el hallazgo más significativo detrás de este AUC es el aumento sustancial en la métrica de *recall* (sensibilidad) por parte del enfoque híbrido cuántico. En las etapas tempranas de descubrimiento de fármacos (*Hit-to-Lead*), los falsos negativos son el peor escenario posible, ya que implican descartar prematuramente un compuesto que podría convertirse en una terapia viable. La codificación a través del *ZZFeatureMap* fue capaz de mapear la geometría del conjunto de datos DYRK1A hacia un hiperespacio cuántico donde moléculas activas que clásicamente parecían "invisibles" o enredadas con el ruido mayoritario, se volvieron linealmente separables.

4.2 Escalabilidad

Para analizar el comportamiento de las métricas a medida que escalaba la dimensión de los datos, simulamos el enfoque QMKL incrementando los qubits (y por tanto los componentes de PCA) desde 4 hasta 13 utilizando un simulador GPU sin ruido. La Tabla 1 resume este comportamiento.

Los datos evidencian que aumentar el espacio de características representadas preserva o amplifica sustancialmente el AUC sobre los modelos clásicos. El rendimiento máximo se alcanza en los **9 qubits** con un AUC de 0.900. No obstante, la simulación de 12 y 13 qubits resalta el severo costo asintótico exponencial de la simulación clásica, incurriendo en 280 y 780 minutos de ejecución respectivamente, demostrando empíricamente por qué una QPU nativa es fundamental a medida que los datos moleculares escalan.

Table 1. Resultados rendimiento cuántico vs. clásico

#qubits	QMKL Performance					Classical performance				
	AUC	Accuracy	Precision	Recall	F1-Score	AUC	Accuracy	Precision	Recall	F1-Score
4	0.867	0.789	0.830	0.830	0.804	0.803	0.711	0.750	0.720	0.735
5	0.873	0.767	0.796	0.780	0.788	0.790	0.722	0.766	0.720	0.742
6	0.873	0.800	0.833	0.800	0.816	0.786	0.744	0.721	0.880	0.793
7	0.871	0.767	0.784	0.800	0.792	0.765	0.700	0.717	0.760	0.738
8	0.895	0.767	0.774	0.820	0.796	0.822	0.767	0.774	0.820	0.796
9	0.900	0.822	0.854	0.820	0.837	0.792	0.711	0.750	0.720	0.735
10	0.897	0.822	0.854	0.820	0.837	0.794	0.744	0.746	0.820	0.781
11	0.891	0.822	0.870	0.800	0.833	0.793	0.756	0.750	0.840	0.793
12	0.884	0.789	0.878	0.720	0.791	0.808	0.744	0.746	0.820	0.781
13	0.871	0.767	0.854	0.700	0.784	0.797	0.756	0.769	0.800	0.784

5 Conclusiones y Trabajo Futuro

Despojando el análisis del optimismo excesivo sobre la disponibilidad a corto plazo de la QRAM, este documento ha evaluado exhaustivamente la viabilidad empírica de las Máquinas de Vectores de Soporte Cuánticas (QSVM) en la era NISQ. Las conclusiones principales dictaminan que el marco de trabajo Q²SAR (basado en *Quantum Multiple Kernel Learning*) no solo ofrece una promesa teórica, sino una ventaja predictiva empíricamente demostrable y aplicable a problemas del mundo real.

En primer lugar, demostramos que la ventaja radica fundamentalmente en la expresividad y la topología del espacio de Hilbert. Frente al desafío matemático de la *dequantization* impulsado por el muestreo de norma L^2 , nuestra investigación justifica que la QSVM no debe aplicarse a matrices de bajo rango. En su lugar, al enfocar el esfuerzo en conjuntos de datos farmacológicos (*datasets* QSAR) donde la correlación de características no lineales es clásicamente intratable, los modelos cuánticos mantienen su separación exponencial frente a los competidores clásicos.

En segundo lugar, hemos abordado y mitigado con éxito el severo cuello de botella de acceso a los datos. La introducción de aceleradores algorítmicos como HoLCUs para la evaluación de valores esperados en un solo circuito representa una innovación no trivial. Al combinarse con *Projected Quantum Kernels* (PQK) para evitar los *Barren Plateaus* y optimizadores primales como Pegasos, la complejidad de ejecución se reduce en factores de hasta 20×, haciendo que los tiempos de las máquinas NISQ sean competitivos.

El beneficio social plausible de este enfoque queda patente en la aplicación al modelado de inhibidores de DYRK1A (una diana para la enfermedad de Alzheimer). Lograr un AUC de 0.8750 frente al 0.8037 de los métodos clásicos demuestra una mejora sustancial que se traduce directamente en la reducción de falsos negativos durante el descubrimiento temprano de fármacos, ahorrando años de investigación y recursos económicos.

Si bien la combinación actual de técnicas hace viable el QSVM hoy en día, el verdadero escalado asintótico exponencial $O(\log(NM))$ prometido teóricamente

requerirá arquitecturas tolerantes a fallos y corrección de errores cuánticos (QEC). Las hojas de ruta de la industria proyectan sistemas con más de 10,000 compuertas lógicas en el futuro cercano. En ese régimen, la inversión directa de matrices mediante subrutinas como el algoritmo HHL, impulsadas por la destilación eficiente de estados mágicos (*Magic States Factories*), permitirá al QSVM procesar bases de datos biomédicas completas sin los factores constantes de penalización por muestreo actuales.

En definitiva, la arquitectura Q²SAR representa una de las herramientas tecnológicas más formidables para empujar las fronteras de la inteligencia artificial científica en la inminente transición hacia la utilidad cuántica práctica.

Acknowledgments. Agradecemos a nuestros colegas y colaboradores por las valiosas discusiones sobre aprendizaje automático cuántico, las cuales fueron fundamentales para el desarrollo de esta investigación. Esta investigación ha sido realizada en el contexto de la Red de Excelencia Cervera para la investigación en tecnologías Cuánticas para Infraestructura eXperimental Orientada a la soberanía Tecnológica en entornos Estratégicos CER – 20251020 | QUIXOTE, financiados por el Ministerio de Ciencia, Innovación y Universidades a través del CDTI.

References

- Burbidge, R., Trotter, M., Buxton, B. F., & Holden, S. B. (2001). Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Computers & Chemistry*, 26(1), 5–14.
- Chia, N.-H., Gilyén, A., Li, T., Lin, H.-H., Tang, E., & Wang, C. (2022). Sampling-based sublinear low-rank matrix arithmetic framework for dequantizing quantum machine learning. *Journal of the ACM (JACM)*, 69(5), 1–72.
- Giraldo, A., Ruiz, D., Caruso, M., et al. (2025). Q²SAR: A quantum multiple kernel learning approach for drug discovery.
- 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, 209–212.
- 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), 2631.
- Mata Ali, A. (2025). Fast expectation value calculation speedup of quantum approximate optimization algorithm: HoLCUs qaoa.
- Pere, C. (2025). Data complexity: A threshold between classical and quantum machine learning - part i.
- Rebentrost, P., Mohseni, M., & Lloyd, S. (2014). Quantum support vector machine for big data classification. *Physical Review Letters*, 113(13), 130503.
- Schuld, M., & Killoran, N. (2019). Quantum machine learning in feature hilbert spaces. *Physical Review Letters*, 122(4), 040504.
- Vedaie, S., Oberoi, A., Zahedinejad, A., et al. (2020). Quantum multiple kernel learning.