JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

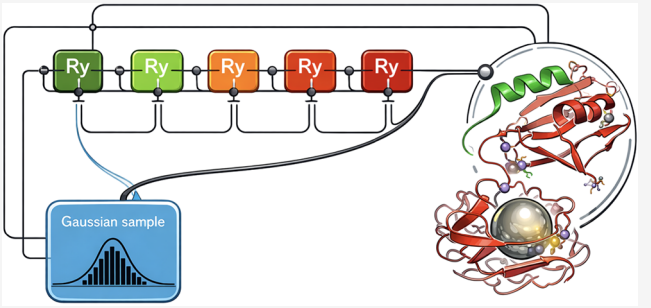
QO-BRA: A Quantum Operator-Based Autoencoder for De Novo Molecular Design

Yue Yu, Francesco Calcagno, Haote Li, and Victor S. Batista*

Cite This: J. Chem. Theory Comput. 2026, 22, 2059−2073 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We introduce a variational quantum autoencoder tailored for de novo molecular design, named **QO-BRA** (Quantum Operator-Based Real Amplitude autoencoder). QO-BRA leverages quantum circuits for real-amplitude encoding and the SWAP test to estimate reconstruction and latent-space regularization errors during back-propagation. Adjoint encoder and decoder operators enable unitary transformations and a generative process that ensures accurate reconstruction, as well as the novelty, uniqueness, and validity of the generated samples. We showcase the capabilities of QO-BRA as applied to the de novo design of Ca^{2+}-, Mg^{2+}-, and Zn^{2+}-binding metalloproteins after training the generative model with a modest data set.



Downloaded via UNIV OF CALIFORNIA BERKELEY on June 1, 2026 at 16:03:28 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

The design of molecular compounds for targeted functions and applications has long been a cornerstone of chemical research.[1,2] With the rise of computational methods, computer-aided molecular design (CAMD) has advanced significantly; however, it continues to face key challenges.[3,4] Early efforts in leveraging structure−function relationships[5,6] have enabled applications ranging from drug delivery to materials science. However, CAMD has remained quite limited due to the complexity of correlating molecular structure with molecular properties in the vast chemical space, which includes a combinatorial number of possible molecules.[7,8]

In recent years, deep learning has driven a new wave of algorithms for molecular design.[9,10] Neural networks (NNs) can now extract complex, hidden patterns from data sets of lead compounds, enabling the generation of novel molecules with structures and properties informed by those in the training set. In fact, popular AI libraries (e.g., DeepChem[11]) are routinely used to predict molecular properties directly from their structure. On the generative side, architectures such as generative adversarial networks (GANs)[12] and reinforcement learning (RL) frameworks[13] can achieve excellent performance in generating valid molecules.

Specifically, deep learning models have focused on protein design.[9,14] Proteins are fundamental to life, carrying out a wide range of functions, including catalysis,[15] transport,[16] signaling,[17] and regulation.[18] They are also implicated in numerous human diseases, such as cancer,[19] diabetes,[20] and Alzheimer's disease,[21] making protein engineering a central challenge in biochemistry. De novo design of proteins thus holds promise for advances in a wide range of applications, including targeted interventions in personalized medicine.[22] It has been shown that neural networks can uncover hidden patterns in natural protein sequences and structures, enabling the generation of artificial proteins with enhanced properties and biologically plausible architectures.[23,24] To date, most models have focused on modifying or improving existing protein scaffolds,[22] while the space of fully de novo protein design remains comparatively less explored.[23] Greener et al.[25] have reported an application of a classical variational autoencoder (VAE) for protein generation, named Protein-VAE, capable of producing novel peptide sequences that bind metal ions by modifying input sequences of up to 140 amino acids. Other powerful models, such as ProteinMPNN[14] and RFdiffusion,[26] have also proven effective in the directed functional design of proteins.

Despite recent advances, classical machine learning models remain limited by computationally intensive encoding schemes, large neural network architectures, extensive training data requirements, and substantial memory footprints. These constraints impede their scalability and complicate efficient tuning for broader applicability. Quantum machine learning (QML) offers a promising alternative, introducing a paradigm shift in computation by exploiting variational quantum circuits that can be trained via back-propagation. QML models can harness the encoding efficiency of quantum superposition

Received: October 11, 2025
Revised: January 23, 2026
Accepted: January 27, 2026
Published: February 3, 2026

ACS Publications © 2026 American Chemical Society

https://doi.org/10.1021/acs.jctc.5c01704
J. Chem. Theory Comput. 2026, 22, 2059−2073

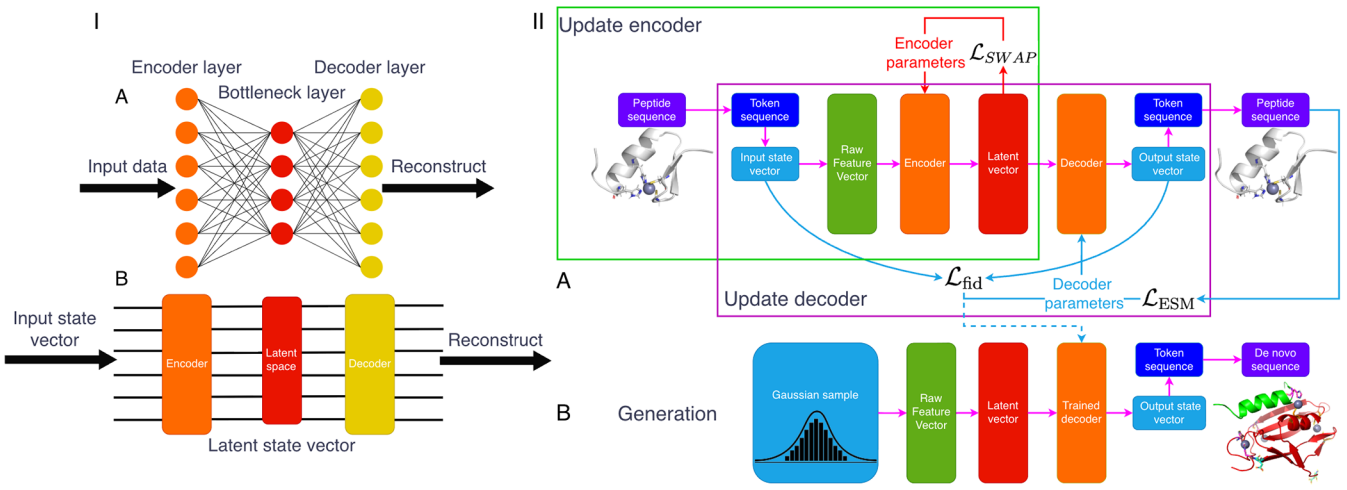


Figure 1. Panel I: Schematic comparison of classical (A) and quantum (B) variational autoencoders. Both architectures include an encoder (orange), a latent space (red), and a decoder (yellow). Panel II: Overview of the QO-BRA model. (A) During training, input peptide sequences are embedded into a quantum circuit (encoder), mapped to a latent space, and reconstructed via the decoder, defined as the adjoint of the encoder. (B) After training, new peptide sequences can be generated by sampling from the learned latent space.

states with inherent parallelism, thereby potentially achieving substantial improvements in computational efficiency.

In the domain of molecular property prediction—including tasks such as toxicity assessment[27,28] and inhibition efficiency estimation[29]—variational quantum models have been investigated as compact, expressive representations for learning chemically relevant features, predominantly within supervised regression and classification paradigms. These studies establish essential baselines for assessing the practical feasibility and performance advantages of QML methods in chemistry. Furthermore, additional research efforts focus on extending variational quantum architectures, such as quantum generative adversarial networks (quantum GANs)[30] and Born machines,[31] to generative modeling and the de novo molecular design problem. However, these approaches are frequently constrained by their dependence on classical neural network components and, in many cases, are only applicable to relatively small molecules, often restricted to systems comprising no more than nine distinct types of heavy atoms.

Superposition states and quantum entanglement should offer key advantages, as they can enable the encoding of correlations that are fundamentally unattainable in classical systems.[32,33] QML models have also demonstrated improved generalization performance and reduced data requirements compared to classical models.[34] Moreover, quantum systems can efficiently represent and manipulate exponentially large state spaces. An N-qubit system encodes 2^N states in parallel; for example, 10 qubits represent 1024 states, while 266 qubits represent approximately 10^{80} states—comparable to the number of atoms in the observable universe.[35] This combination of exponentially scalable state representation and lower data demands positions QML as a promising approach for domains such as molecular design, where combinatorial complexity and limited training data present major bottlenecks.

Quantum variational autoencoders (QVAEs) are emerging as powerful tools for processing quantum data and simulating quantum systems. These models combine classical variational autoencoders with quantum components to enable efficient compression, representation learning, and generation of quantum states.[36,37] QVAEs have demonstrated competitive performance on tasks such as image generation and can be trained using quantum Monte Carlo simulations.[36] Recent advancements include the ζ-QVAE, which utilizes regularized mixed-state latent representations and can be applied directly to quantum data.[37] Additionally, quantum circuit autoencoders have been developed to compress information within quantum circuits, with applications in anomaly detection and noise mitigation.[38] These quantum autoencoder models show promise in learning efficient representations of quantum states, including those that are difficult to simulate classically, suggesting potential applications in near-term quantum hardware.[39]

Here, we introduce a QVAE tailored for de novo molecular design autoencoder named QO-BRA (Quantum Operator-Based Real Amplitude), schematically illustrated in Figure 1IB. QO-BRA is a generative model that learns to encode input data into a continuous, low-dimensional latent space and decode it to reconstruct the original data. Unlike conventional autoencoders, VAEs impose a probabilistic structure—typically a multivariate Gaussian—on the latent space. This regularization enables smooth interpolation between latent representations and the conditional generation of molecules in close chemical proximity to a reference structure.[10,25,40] When appropriately trained, VAEs can generate novel compounds that preserve key characteristics of the training distribution. Prior work has demonstrated their utility across a range of molecular design tasks, including the generation of molecules with tailored physicochemical properties, selective binding affinities, or compatibility with specific retrosynthetic routes,[10,40,41] as well as applications in protein design[23,25] and molecular structure prediction.[12,42] QO-BRA is agnostic to the specific quantum computing platform; therefore, we describe how to implement it on conventional qubit-based devices (Part I) as well as on hybrid qubit-qumode platforms.[43]

From an architectural perspective, QO-BRA departs from existing quantum variational autoencoder (QVAE) frameworks in several key respects. Whereas prior QVAEs typically rely on independently parametrized encoder—decoder circuits and overlap-based or measurement-intensive loss functions,[36,37,44] QO-BRA employs a SWAP-test—derived fidelity objective to quantify reconstruction quality directly in amplitude space. This approach provides a principled measure of state similarity

https://doi.org/10.1021/acs.jctc.5c01704
J. Chem. Theory Comput. 2026, 22, 2059−2073

Table 1. Qualitative Comparison between Representative Classical Protein Design Models and QO-BRA, Illustrating Differences in Modeling Paradigms, Training Scale, and Intended Use Cases

Model	Protein-VAE[25]	ProteinMPNN[14]	RFdiffusion[26]	QO-BRA
Primary Task	Conditional protein sequence generation (metal binding, fold grammar)	Structure-conditioned sequence design	Generative fold and functional-site design	Latent-space generative molecular/protein design
Model Paradigm	Classical conditional variational autoencoder	Graph neural network	Diffusion model with SE(3)-equivariant networks	Variational quantum autoencoder
Structural Conditioning	Grammar-based topology encoding; metal-binding labels	Explicit backbone geometry	Explicit structural and motif constraints	Implicit via encoded molecular/protein features
Primary Strengths	Interpretable latent space; Conditional generation	High sequence recovery accuracy	Flexible generation of novel folds and binders	Data efficiency; Compact latent representation
Training Data Scale	$\sim 10^5$ sequences (including homologues)	$\sim 10^5$–10^6 protein structures	$\sim 10^5$–10^6 protein structures	$\sim 6 \times 10^3$ sequences
Parameter Count	$\sim 10^5$–10^6 classical parameters	$\sim 10^5$–10^6 classical parameters	$\sim 10^6$–10^7 classical parameters	18–27 variational quantum parameters

while avoiding full quantum state tomography and reducing measurement overhead in simulation-based training.

A defining feature of QO-BRA is its adjoint decoder architecture, in which the decoding operation is implemented as the inverse of the encoder circuit. This design enforces information-theoretic consistency between encoding and decoding, substantially reduces the number of trainable variational parameters, and enables stable generative decoding in shallow-circuit regimes. To our knowledge, this adjoint construction has not been explored in prior QVAE-based molecular design models.

The encoder circuit is implemented using a RealAmplitudes ansätze,[45] which restricts the accessible unitary space relative to the full $SU(2^n)$ manifold. This constraint is imposed deliberately to promote optimization stability in data-limited regimes, preserve a structured and interpretable latent representation, and maintain circuit depths compatible with NISQ-scale evaluation. As a result, QO-BRA can be trained and evaluated consistently across noisy simulators, NISQ hardware, and high-performance state vector backends, providing a concrete realization of platform-agnostic quantum generative modeling.

We illustrate QO-BRA as applied to de novo protein design. Hence, we demonstrate the effectiveness of QO-BRA in generating metalloproteins that selectively bind divalent metal ions, including Ca^{2+}, Mg^{2+}, and Zn^{2+}. The model reliably produces appropriate metal-binding sites as defined by both the primary amino acid sequence and the spatial arrangement of coordinating side chains. QO-BRA exhibits strong robustness to hyperparameter variation and consistently delivers high-quality designs using minimal training data and a compact set of variational parameters.

The aforementioned classical models are each grounded in distinct computational paradigms. ProteinMPNN, for example, is a state-of-the-art supervised model for protein sequence design, whereas RFdiffusion belongs to a class of diffusion-based generative frameworks for protein fold and binder design. These approaches typically operate in structure-to-sequence or structure-conditioned generative settings, rely on large-scale training data sets, and employ models with millions to billions of parameters to achieve high-resolution sequence recovery.

In contrast, QO-BRA emphasizes amplitude-space latent encodings, is designed for low-data or parameter-constrained regimes, and investigates quantum circuit–based generative models that are fundamentally distinct from classical architectures in their representational mechanisms. QO-BRA is not intended to surpass or supplant state-of-the-art classical models; rather, its objective is to assess the feasibility of quantum-enhanced latent spaces, to function effectively under conditions of limited training data or restricted model capacity, and to evaluate generative reconstruction fidelity on molecular fragments and fold motifs considered as modular building blocks.

To place QO-BRA in context, Table 1 summarizes representative classical generative protein design paradigms alongside our quantum variational autoencoder formulation. In contrast to large-scale diffusion- and graph-based pipelines that typically rely on extensive training corpora and high-capacity neural architectures, QO-BRA is intentionally developed for a parameter- and data-limited regime, enabling latent-space sampling and reconstruction under constrained model capacity. Accordingly, QO-BRA is not positioned as a direct competitor to state-of-the-art classical design workflows; rather, it serves as an exploratory platform for assessing quantum variational representations and their potential role within future hybrid quantum–classical protein design strategies.

RESULTS AND DISCUSSION

This section presents the results of QO-BRA-driven de novo generation of Ca^{2+}-, Mg^{2+}-, and Zn^{2+}-binding proteins. We begin by analyzing the impact of key hyperparameters on generation performance, with a particular focus on the ansätze unit repetition number (r) and the number of qubits (N_q). Their influence on both model efficiency and structural quality is systematically investigated. We then highlight representative metalloproteins generated using the optimal hyperparameter configurations, demonstrating that QO-BRA produces high-quality protein structures.

Generation quality is assessed by comparing the features of the generated proteins against those in the training set. Specifically, we examine token frequency distributions, peptide length distributions, the number of ion binding sites, and the number of chains per complex. To quantify the alignment between the generated and training data, we compute a relative ratio (RR) for each of these four properties. An ideal model would yield $RR = 1$ across all metrics. In addition, we evaluate the generated sequences using the NUVR metric, which assesses novelty (N), uniqueness (U), validity (V), and reconstruction accuracy (R). Each component is scored between 0 and 1, with 1 indicating a sequence that is entirely novel, unique, chemically valid, and accurately reconstructed. Further methodological details are provided in the Supporting Information.[40]

https://doi.org/10.1021/acs.jctc.5c01704
J. Chem. Theory Comput. 2026, 22, 2059–2073

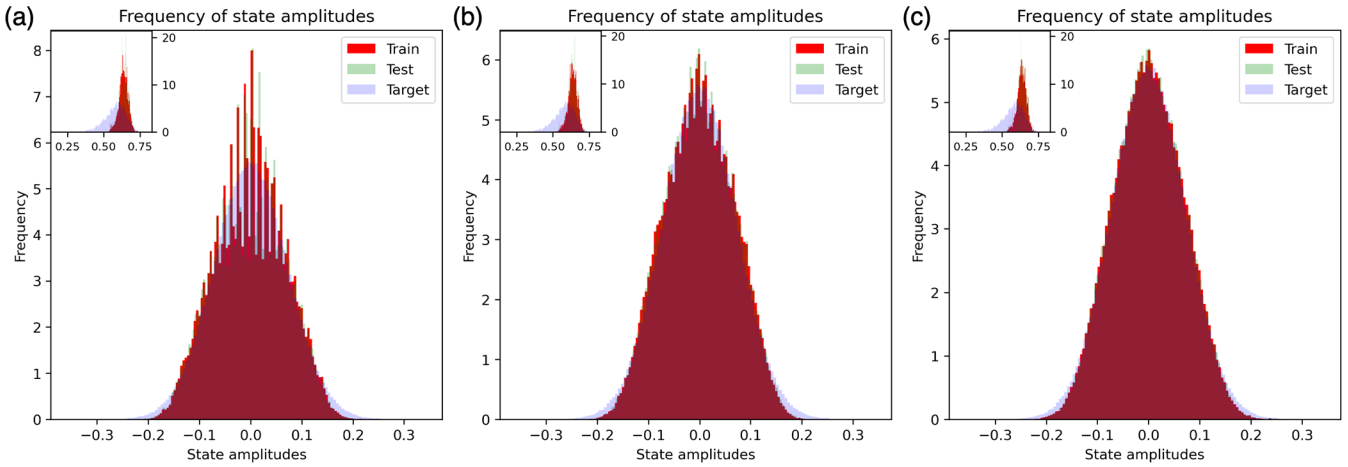


Figure 2. Latent space fitting after training on Zn^{2+} data with $N_q = 7$ for different ansätze depths: $r = 1$ (a), $r = 2$ (b), and $r = 3$ (c), along with how long the training takes on emulations of the corresponding quantum circuits. Increased depth leads to improved alignment with the target distribution, reflecting higher model expressivity. This improvement is most evident from $r = 1$ to $r = 2$, and marginally from $r = 2$ to $r = 3$. Provided the trade-off between network expressivity and complexity, subsequent QO-BRA operations are done with $r = 2$.

Quantum Network Structure Effects

Effect of Ansätze Depth (r) on Model Performance. In classical convolutional neural networks, model capacity is strongly influenced by both depth and the number of trainable parameters.[46] Analogously, in QML, circuit depth plays a critical role in model expressivity and learning performance. In QO-BRA, this depth is governed by the number of repetitions r of the RA ansätze.

Figure 2 shows the latent space fitting quality after training QO-BRA with $r \in \{1, 2, 3\}$. The inset highlights the first component of the latent vectors, illustrating how the "head" of the sequence is embedded in latent space. The main plots display the fitting behavior of the remaining components. The target latent distribution is a Gaussian with zero mean and standard deviation $\sigma = (1.5 \times 2^{N_q/2})^{-1}$. For $r = 1$ (Figure 2a), the model shows a limited ability to match the target distribution. Increasing to $r = 2$ (Figure 2b) significantly improves the fit, indicating that a deeper ansätze enhances learning capacity. A further increase to $r = 3$ (Figure 2c) offers only marginal improvements, suggesting that additional depth yields diminishing returns.

As detailed in Table 2, increasing r leads to a linear growth in the number of trainable parameters and a corresponding increase in training time on an emulation of the quantum circuits. Based on this trade-off between performance and efficiency, we fix $r = 2$ for all subsequent experiments.

Table 2. Encoder Parameter Count and Training Runtime for Zn^{2+} Data as a Function of Ansätze Depth r, with $N_q = 7$[a]

r	Parameters	Training Runtime (h)
1	14	3.94
2	21	5.90
3	28	7.03

[a]Training was performed using 48 x86_64 Intel CPUs. Only encoder parameters are reported.

Trade-Off between Qubit Count (N_q) and Model Capacity. Another key hyperparameter is the total number of qubits, N_q, which defines the maximum peptide length that the model can process. Specifically, a network with N_q qubits can handle sequences of up to $2^{N_q} - 1$ residues. If N_q is too small, the model cannot generate sufficiently long or complex sequences to represent functional proteins. On the other hand, while the theoretical advantage of QML partly stems from scaling with the qubit number,[47] increasing N_q leads to a linear growth in the number of trainable parameters. This significantly increases the computational cost and training time. To balance expressivity and efficiency, we restrict our exploration to $N_q = 6$, 7, 8, and 9, as shown in Tabs. 3 and 4.

While the NUVR metric remains relatively consistent across the three ion data sets (Table 3), the relative ratio (RR) results—summarized in Table 4—highlight a more nuanced dependence on the qubit count N_q. In general, performance improves with increasing N_q, as reflected by RR values approaching the ideal value of 1 across all training scenarios. This trend is most pronounced for Zn^{2+} at $N_q = 9$, as shown in Figure 3, where the generated sequences closely match the training distribution across all evaluated metrics: token frequency, number of chains, complex size, and number of binding sites. A broader analysis across all three ions reinforces this pattern. For both Ca^{2+} and Zn^{2+}, the RR values consistently converge toward 1 as N_q increases—ranging from 0.51 ± 0.68 to 5.05 ± 11.28 for Ca^{2+}, and from 1.01 ± 1.39 to 4.52 ± 5.72 for Zn^{2+}. Although Mg^{2+} exhibits greater variance and less favorable alignment with the training distribution (RR range: 2.58 ± 2.20 to 10.07 ± 26.03), the underlying trend of improved distributional similarity with increasing N_q remains consistent. Based on this observation, we fix $N_q = 9$ in all subsequent experiments, enabling the model to generate primary sequences of up to 511 amino acids.

Tertiary Structure Prediction and Refinement

In de novo metalloprotein design, accurate reconstruction of tertiary structure from a generated primary sequence is essential for assessing functional viability—particularly for identifying and localizing metal ion binding sites. To enable this, we implemented a structure prediction pipeline tailored to QO-BRA-generated sequences (Figure 4).

Sequence-to-Structure Workflow. Figure 4 illustrates the computational pipeline used to convert a generated primary sequence into a fully solvated, structurally equilibrated

https://doi.org/10.1021/acs.jctc.5c01704
J. Chem. Theory Comput. 2026, 22, 2059−2073

Table 3. NUVR Metric Components—Novelty (N), Uniqueness (U), Validity (V), and Reconstruction Accuracy (R)—for Generated Sequences, Evaluated on Training and Test Sets[a]

Ion, N_q	Parameters	N	U	V	R_{train}	R_{test}	$NUVR_{train}$
Ca^{2+}, 6	18	0.98	1.00	0.86	1.00	1.00	0.84
Ca^{2+}, 7	21	0.92	1.00	0.85	1.00	1.00	0.79
Ca^{2+}, 8	24	0.91	1.00	0.82	1.00	1.00	0.75
Ca^{2+}, 9	27	0.93	1.00	0.84	1.00	1.00	0.78
Ca^{2+}, CNN-VAE	$>9.4 \times 10^6$	1.00	0.60	0.60	0.17	0.11	0.062
Mg^{2+}, 6	18	0.99	1.00	0.80	1.00	1.00	0.79
Mg^{2+}, 7	21	0.97	1.00	0.83	1.00	1.00	0.81
Mg^{2+}, 8	24	0.96	1.00	0.76	1.00	1.00	0.73
Mg^{2+}, 9	27	0.96	1.00	0.76	1.00	1.00	0.73
Mg^{2+}, CNN-VAE	$>9.4 \times 10^6$	1.00	0.65	0.65	0.085	0.048	0.036
Zn^{2+}, 6	18	0.84	1.00	0.81	1.00	1.00	0.68
Zn^{2+}, 7	21	0.81	1.00	0.84	1.00	1.00	0.68
Zn^{2+}, 8	24	0.78	1.00	0.80	1.00	1.00	0.62
Zn^{2+}, 9	27	0.86	1.00	0.80	1.00	1.00	0.69
Zn^{2+}, CNN-VAE	$>9.4 \times 10^6$	0.98	0.50	0.50	0.16	0.090	0.040

[a]Results are shown for each ion type and qubit count N_q, as well as for the classical CNN-based VAE for comparison. The composite $NUVR_{train}$ score reflects generation quality under each configuration.

Table 4. Relative Ratio (RR) Metrics for Token Frequency, Number of Chains, Peptide Length, and Binding Sites are Computed across Different Ion Types and Qubit Counts (N_q), Including Measurements from the Classical CNN-Based VAE, Which Takes in a Maximum $2^9 = 512$ Token Length, Corresponding to $N_q = 9$[a]

Ion, N_q	Parameters	Token Freq.	Chains	Length	Binding Sites
Ca^{2+}, 6	18	1.69 ± 2.01	2.75 ± 3.42	9.85 ± 8.21	4.89 ± 8.35
Ca^{2+}, 7	21	2.13 ± 2.23	7.62 ± 8.58	23.71 ± 27.47	3.24 ± 3.85
Ca^{2+}, 8	24	1.82 ± 1.31	9.91 ± 13.06	13.34 ± 23.09	1.12 ± 0.94
Ca^{2+}, 9	27	1.12 ± 0.62	1.05 ± 0.60	5.05 ± 11.28	0.51 ± 0.68
Ca^{2+}, CNN-VAE	$>9.4 \times 10^6$	0.88 ± 0.18	0.89 ± 0.84	3.41 ± 7.52	0.59 ± 0.69
Mg^{2+}, 6	18	4.15 ± 9.20	16.34 ± 14.80	23.12 ± 22.18	27.35 ± 37.99
Mg^{2+}, 7	21	6.06 ± 12.01	30.95 ± 36.58	46.94 ± 47.87	46.26 ± 47.54
Mg^{2+}, 8	24	5.04 ± 5.77	30.38 ± 48.18	23.09 ± 43.33	21.71 ± 26.31
Mg^{2+}, 9	27	2.58 ± 2.20	4.07 ± 4.01	10.07 ± 26.03	4.55 ± 4.04
Mg^{2+}, CNN-VAE	$>9.4 \times 10^6$	0.96 ± 0.16	0.87 ± 0.78	3.41 ± 6.60	0.56 ± 0.72
Zn^{2+}, 6	18	14.43 ± 40.71	2.37 ± 2.91	3.75 ± 2.13	7.72 ± 11.48
Zn^{2+}, 7	21	17.36 ± 40.76	3.40 ± 5.23	4.24 ± 1.89	5.25 ± 5.22
Zn^{2+}, 8	24	11.35 ± 14.92	2.68 ± 5.40	3.36 ± 2.26	3.38 ± 4.31
Zn^{2+}, 9	27	4.52 ± 5.72	1.19 ± 1.51	1.73 ± 1.37	1.01 ± 1.39
Zn^{2+}, CNN-VAE	$>9.4 \times 10^6$	0.90 ± 0.31	1.23 ± 1.00	1.85 ± 1.83	1.38 ± 1.33
Zn^{2+}, Random	N/A	0.91 ± 0.23	0.66 ± 1.58	1.06 ± 0.31	0.76 ± 1.25

[a]Moreover, for Zn^{2+}, the metric values for random generation are also put in contrast. Each row also lists the total number of encoder parameters. Higher N_q allows longer sequences but increases model complexity.

protein model suitable for downstream analysis. The workflow consists of four main stages:

1. **Sequence Input and Formatting:** The pipeline begins with a peptide sequence generated by QO-BRA, stored in a plain text file (`sequence.txt`). This sequence is converted into standard `FASTA` format to ensure compatibility with structure prediction tools.
2. **Structure Prediction (Chai-1):** The FASTA file is processed by the Chai-1 structure prediction engine,[48] which outputs a predicted 3D conformation in `.cif` format. This file is then converted to `PDB` format, representing the protein atomic coordinates in the absence of solvent and ions—referred to as the *dry PDB*.
3. **Solvation and Molecular Dynamics Simulation (OpenMM 8):** The dry structure is input into OpenMM 8,[49] where it is solvated using the TIP3P water model and neutralized with counterions. A molecular dynamics (MD) simulation is then performed to equilibrate the structure under near-physiological conditions. The resulting output is an equilibrated *aqueous PDB* that incorporates solvent and ion coordination effects.
4. **Structural Analysis:** The equilibrated structure is subsequently subjected to structural analysis, including RMSD calculations and evaluation of binding site integrity. These measurements provide insight into the stability of the de novo generated protein models.

This modular workflow enables the reliable translation of synthetic sequences into realistic 3D structures for functional and biophysical characterization.

Three-Dimensional Protein Structures. Three-dimensional structure prediction was performed using Chai-1,[48] a state-of-the-art deep learning framework for modeling protein conformations. Representative outputs of Chai-1 applied to QO-BRA-generated sequences are shown in Figure 5. This task

https://doi.org/10.1021/acs.jctc.5c01704
J. Chem. Theory Comput. 2026, 22, 2059−2073

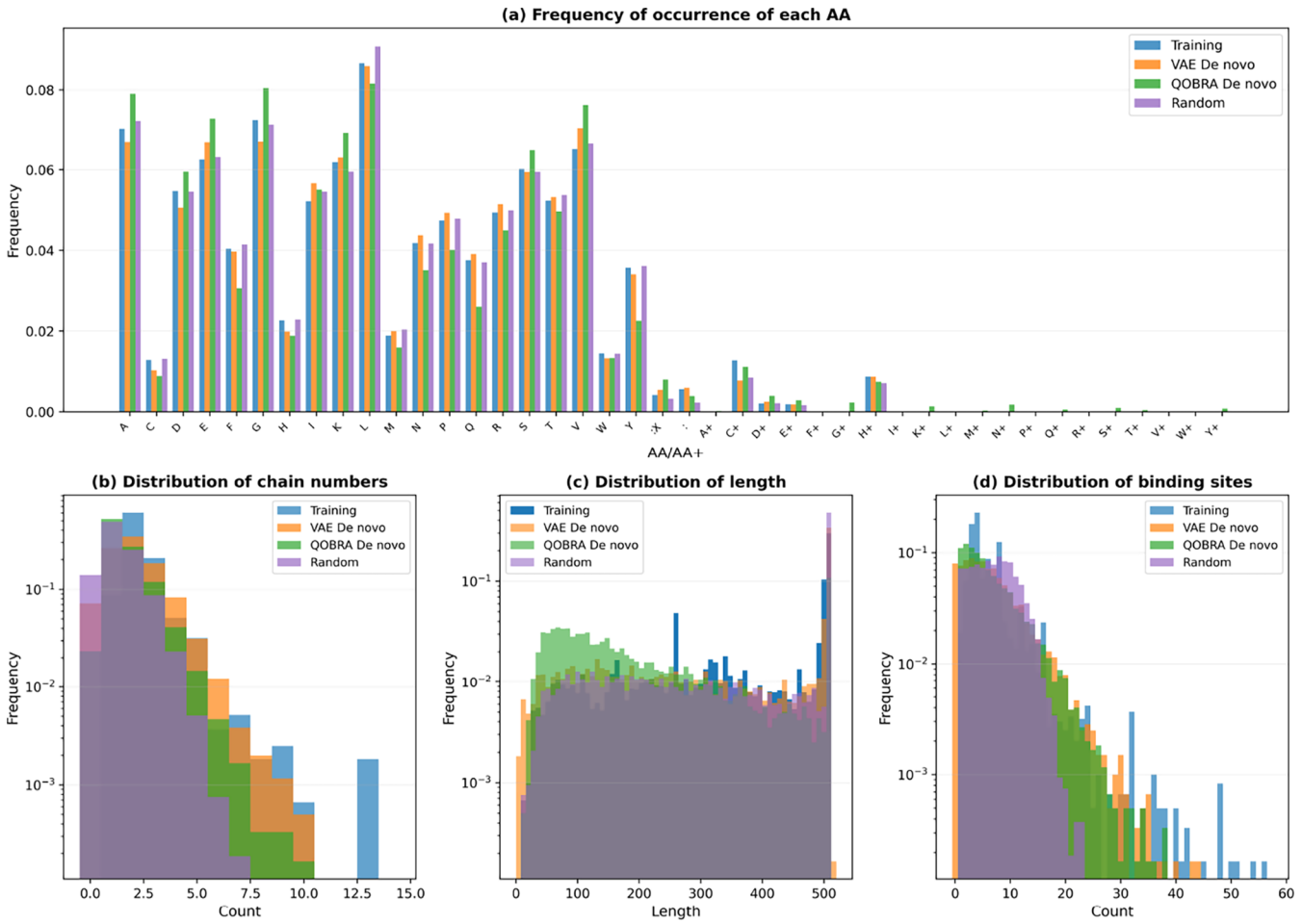


Figure 3. Histograms for Zn^{2+} with N_q = 9 of the frequencies of tokens (a), chain numbers (b), peptide lengths (c), and ion binding sites (d) comparing classical CNN-generated (orange) and QO-BRA-generated sequences (green) to the training set (blue). A comparison is also made against instances produced by randomly sampling token based on their frequencies in the training set (purple). The length is calculated as the number of AAs plus: in a sequence, while chain number is computed as how many: a sequence contains. A 0 chain number implies that the sequence is a partial domain within a larger complex.

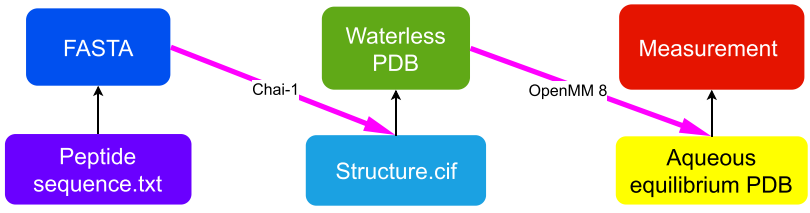


Figure 4. Schematic overview of the sequence-to-structure pipeline. A generated primary sequence is formatted and converted to FASTA, processed by the Chai-1 language model to predict a three-dimensional structure (in CIF and PDBformats), and subsequently equilibrated via molecular dynamics simulation in OpenMM 8 to produce a solvated, biologically relevant structure.

presents a nontrivial challenge: the generated sequences are synthetic and lack homologues in structural databases, precluding the use of homology-based modeling. Consequently, Chai-1 infers structural configurations in a purely *ab initio* manner. Metal ion placement is handled iteratively, with ions introduced into the structure until all predicted coordination sites are saturated based on local residue geometry.

To ensure structural stability, all predicted conformations were subjected to molecular dynamics (MD) refinement in explicit solvent. Simulations were carried out using OpenMM 8[49] at a constant temperature of 300 K. Protein interactions were described using the AMBER14 force field,[50] while the solvent was modeled using the TIP3P water model.[51] Each structure was solvated in a cubic water box extending 0.5 nm beyond the protein in all dimensions, and counterions (Na^+, Cl^-) were added to neutralize the net charge.

Systems underwent energy minimization using Langevin dynamics for 50,000 steps, followed by temperature equilibration to 300 K via a Langevin thermostat,[52] employing a 4 fs integration time step and a friction coefficient of 1 ps^{-1} over an additional 50,000 steps. Structural stability and convergence were assessed throughout the simulation using root-mean-square deviation (RMSD) analysis, calculated with MDTraj.[53] This refinement pipeline produces solvent-equilibrated struc-

https://doi.org/10.1021/acs.jctc.5c01704
J. Chem. Theory Comput. 2026, 22, 2059−2073

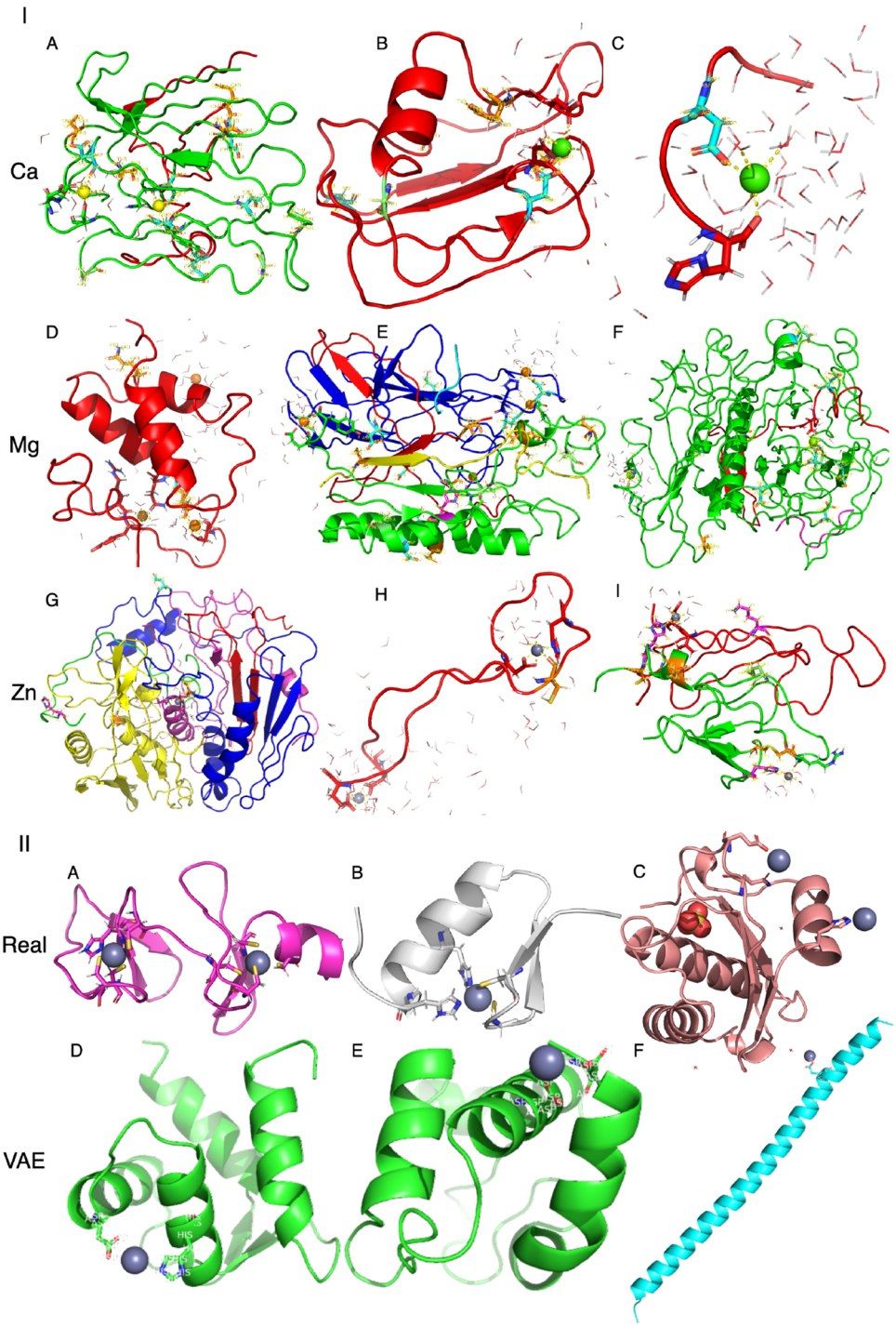


Figure 5. (A) Representative artificial metalloproteins generated by QO-BRA with $N_q = 9$ and $r = 2$. Structures include Ca^{2+}-binding (green, A1−A3), Mg^{2+}-binding (lime, A4−A6), and Zn^{2+}-binding (gray, A7−A9) proteins. Tertiary structures were predicted using Chai-1.[48] Highlighted residues indicate predicted ion-coordinating sites identified by the QO-BRA model. Coordinating water molecules are also shown, forming metal-specific coordination geometries−hexahedral for Ca^{2+} and Mg^{2+}, tetrahedral for Zn^{2+}. (B) Examples of Zn^{2+}-binding proteins from nature (B1−B3) and from sequences generated by the Protein-VAE model of Greener et al. (B4−B6).

tures, allowing direct comparison to natural metalloproteins and enabling downstream biophysical or functional analysis.

Selectivity and Specificity. The primary sequences generated by QO-BRA contain canonical secondary structure elements, including α-helices, β-sheets, and coils−in proportions comparable to those observed in the training set (α-helices: 30−45%; β-sheets: 20−30%; loops/turns/other: 25−40%). These sequences fold into tertiary structures that closely resemble those of natural proteins, as illustrated by representative examples in Figure 5A. Furthermore, the predicted metal-binding sites agree with established principles of coordination chemistry, with preferred ligands being the amino acid side chains that are distinct for each type of metal.

We define a Chai-1 prediction as successful if the predicted 3D structure places a metal ion in close proximity to the residues identified by QO-BRA as metal-coordinating. False

https://doi.org/10.1021/acs.jctc.5c01704
J. Chem. Theory Comput. 2026, 22, 2059−2073

positives (FP) occur when predicted coordinating residues lack nearby metal ions, while false negatives (FN) are residues not predicted by QO-BRA but located near metal ions in the structure. True positives (TP) and true negatives (TN) follow the standard definitions. From these, we compute sensitivity and specificity as follows

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{specificity} = 1 - \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{1}$$

We have evaluated 100 generated structures per metalloprotein type. Coordination was assessed using metal-specific cutoff distances, identifying coordinating atoms from side chains or water molecules. Histogram distributions of sensitivity and specificity are shown in Figure S1. Overall, the model achieves high specificity, with most values in the (0.9, 1.0) range across all three ions. Sensitivity, however, is more variable, often peaking near zero, indicating missed coordinating residues. Nonetheless, occasional cases of 100% sensitivity demonstrate that the model is capable of high performance under the right structural conditions.

Simulations involving Zn^{2+} consistently show coordination pockets composed of residues known to biologically bind Zn^{2+}—histidine, cysteine, aspartate, and glutamate—along with water molecules. These tertiary motifs, consistent with natural and engineered proteins,[54,55] also emerge in QO-BRA-derived structures. Similar trends are observed for Ca^{2+} and Mg^{2+}, which preferentially coordinate with aspartate, glutamate, and water.[56,57] The predicted binding pockets typically include both QO-BRA-predicted residues and additional structural contributors to the coordination sphere.

Coordination Number. A more rigorous assessment of the structural quality of the generated protein models can be obtained by analyzing the coordination number of the bound metal ions—that is, the number of atoms directly coordinating each ion. Coordination numbers are ion-specific and are influenced by both the identity of the ion and the nature of its ligands, including water and nonpeptidic molecules.[58,59]

In aqueous protein environments, calcium (Ca^{2+}) typically adopts coordination numbers of 6 to 8, magnesium (Mg^{2+}) commonly coordinates with 6 atoms, and zinc (Zn^{2+}) generally exhibits coordination numbers between 4 and 6. Each ion also has characteristic coordination distances that reflect its size and preferred ligand geometries.

Figure S2 presents the coordination numbers and corresponding distances observed in our generated structures in contrast to those observed for the authentic metalloproteins included in the training set. Overall, the resulting distributions exhibit good agreement, with the frequencies associated with each coordination number displaying comparable trends between the experimentally derived and the computationally generated data.

Secondary Structure Proportions. Table 5 illustrates the proportions of the three secondary structures of proteins for each ion set compared to the expected range for natural proteins. The measurements were performed using DSSP[60] in Biopython,[61] which provides results closely aligned with natural ranges, notwithstanding the tendency of protein language models such as Chai-1 to predict helical structures.[48,62,63] This suggests that QO-BRA possesses a degree of capability in understanding the primary sequence composition of proteins to create proxy-natural proportions of domains.

Table 5. Protein Secondary Structure Proportions in Three Types of Generated Structural Sets vs Proportions in Natural Proteins

	α-helix	β-sheet	Coil
Ca^{2+}	0.30 ± 0.20	0.24 ± 0.15	0.46 ± 0.14
Mg^{2+}	0.34 ± 0.20	0.21 ± 0.15	0.45 ± 0.15
Zn^{2+}	0.34 ± 0.20	0.20 ± 0.15	0.46 ± 0.12
Natural	[0.3, 0.35]	[0.2, 0.25]	[0.4, 0.5]

Primary Sequence Analysis

Normalized DOPE score. To evaluate the intrinsic quality and physical plausibility of the de novo generated protein folds, we employ the Discrete Optimized Protein Energy (DOPE) score,[64] a knowledge-based statistical potential that depends on interatomic distances. More negative DOPE values, which correspond to lower estimated free energy, are indicative of higher-quality, more probable native-like conformations. However, the raw DOPE score is influenced by protein length and amino acid composition, rendering it suboptimal for direct comparison across structurally diverse proteins. To address this limitation, Eramian et al.[65] introduced a normalized variant, N-DOPE, which rescales the DOPE score to facilitate meaningful cross-structure comparisons. In their work, N-DOPE values below −1.5 are reported to be characteristic of near-native models, whereas values above 1.0 are typically associated with physically implausible structures.

In Figure S3, the N-DOPE scores are reported for Ca^{2+}—(a–c), Mg^{2+}—(d–f), and Zn^{2+}-binding proteins (g–i), with structures comprising, from left to right, one to three coordinated ions. Across all sets, models classified as invalid constitute only a small fraction within the invalid-score region and are approximately evenly partitioned between near-native and ambiguous conformations. Taken together, these observations indicate that QO-BRA exhibits a substantive capability to generate sequence-derived tertiary structures that are structurally plausible.

Natural vs Generated Sequences. The synergy between the generative capabilities of QO-BRA and the structure prediction provided by Chai-1 demonstrates an effective approach to designing protein sequences and structures. Both models recover fundamental biophysical patterns and generate novel proteins that closely replicate the composition and architecture of natural systems. This level of performance is especially notable given the minimal parameter count—only 27 trainable variables—and the modest training set size of approximately 6,000 sequences per metal ion. In comparison, a generative classical model employed a conventional variational autoencoder with 912 neurons, four hidden layers, and more than 105,000 sequences to achieve similar results (Figure 5B).[25] These outcomes are made possible by the distinctive architecture and operational principles of QO-BRA, which differ substantially from those of classical machine learning methods.

Baseline Models for Contextualizing QO-BRA. To contextualize the performance metrics reported in Tables 3 and 4, we considered two reference baselines: (i) a conventional convolutional neural network–based variational autoencoder (CNN-VAE) implemented in PyTorch Lightning,[66,67] and (ii) a random-sampling control. The CNN-VAE employs one-dimensional convolutional layers to model local sequence motifs and short-range dependencies, and it was trained on the same data sets as QO-BRA using a fixed input

https://doi.org/10.1021/acs.jctc.5c01704
J. Chem. Theory Comput. 2026, 22, 2059−2073

Table 6. Pairwise Distances between ESM2-Derived Mean Per-Residue Log-Likelihood Distributions (Zn^{2+}, N_q = 9) for the Training Set, Random Sampling, the Classical CNN-VAE, and QO-BRA[a]

KS Statistic				
	Training Set	Random	CNN-VAE	QO-BRA
Training Set	0.000	0.9749	0.9505	0.5785
Random	0.9749	0.000	0.4551	0.9998
CNN-VAE	0.9505	0.4551	0.000	0.9928
QO-BRA	0.5785	0.9998	0.9928	0.000
KS *p*-value				
	Training set	Random	CNN-VAE	QO-BRA
Training Set	1.00	0.00	0.00	1.00×10^{-6}
Random	0.00	1.00	0.00	3.13×10^{-55}
CNN-VAE	0.00	0.00	1.00	1.73×10^{-41}
QO-BRA	1.00×10^{-6}	3.13×10^{-55}	1.73×10^{-41}	1.00
Wasserstein Distance				
	Training Set	Random	CNN-VAE	QO-BRA
Training Set	0.000	0.3120	0.2973	0.1691
Random	0.3120	0.000	0.0150	0.1459
CNN-VAE	0.2973	0.0150	0.000	0.1312
QO-BRA	0.1691	0.1459	0.1312	0.000

[a]We report the two-sample Kolmogorov–Smirnov (KS) statistic[69] with its associated *p*-value, and the 1D Wasserstein distance.[70]

length of 2^9 = 512 tokens, corresponding to the maximum sequence length that can be encoded by a quantum circuit with N_q = 9 qubits.

For the Zn^{2+} data set, the random baseline was constructed by independently sampling amino acid tokens according to their empirical marginal frequencies estimated from the Zn^{2+} training set. This procedure preserves single-residue composition while eliminating higher-order correlations and sequence context, thereby yielding a composition-matched null model that facilitates the separation of learned sequence-organization effects from those driven purely by token-frequency distributions.

As summarized in Table 3, QO-BRA attains higher composite NUVR scores than the CNN-based VAE while employing only 27 trainable parameters, in contrast to more than 9.4×10^6 trainable parameters in the classical architecture. This pronounced separation in performance is primarily attributable to the consistently high reconstruction accuracy (R) achieved by QO-BRA across ionic species and qubit configurations, whereas the classical model exhibits appreciably reduced reconstruction fidelity. Consequently, these observations delineate distinct operational regimes for the two approaches, rather than constituting a direct state-of-the-art benchmarking comparison. Complementarily, the RR metrics reported in Table 4 indicate that, for Zn^{2+}, QO-BRA and the CNN-VAE yield broadly comparable distributional behavior across multiple sequence-level statistics, whereas random token sampling deviates substantially from both methods over several of these metrics.

To evaluate the extent to which the generated sequences preserve training-like residue-level statistics, we compared the empirical distributions of mean per-residue log-likelihoods derived from Evolutionary Scale Modeling (ESM2), a pretrained transformer-based protein language model.[68] The details are in subsection ESM2 Evaluation. We applied it to the training set and three generated ensembles (Figure S4). Pairwise distributional similarity was quantified using the two-sample Kolmogorov–Smirnov (KS) statistic[69] (with corresponding *p*-values) and the one-dimensional Wasserstein distance[70] (Table 6). Among the generative baselines, QO-BRA exhibits the closest agreement with the training distribution, with W(Training set, QO-BRA) = 0.1691 compared to W(Training set, CNN–VAE) = 0.2973 and W(Training set, Random) = 0.3120, and likewise KS(Training set, QO-BRA) = 0.5785 versus KS(Training set, CNN–VAE) = 0.9505 and KS(Training set, Random) = 0.9749. Collectively, these results indicate that, under this diagnostic, QO-BRA most closely recapitulates the training-set likelihood landscape as inferred by ESM2.

MATERIALS AND METHODS

An overview of the QO-BRA workflow is shown in Figure 1IIA. The architecture consists of two components: a *quantum encoder* and a *quantum decoder*. Both are implemented as parametrized quantum circuits, mirroring the structure of classical neural networks (cNNs). The circuit variational parameters are optimized during training by back-propagation.

In our applications for de novo protein design, input peptide sequences are mapped into quantum amplitudes through a letter-to-number encoding scheme, followed by normalization. This transforms discrete token sequences into continuous quantum state vectors suitable for processing by the quantum encoder.

Training jointly optimizes two loss functions in a self-consistent loop, ensuring regularization into a Gaussian latent space distribution and accurate reconstruction through direct comparisons with SWAP tests (detailed in Loss Functions).

Following training, the decoder operates independently (Figure 1IIB) to generate novel peptide sequences. This is achieved by sampling from the latent space, applying the decoder gates to the sampled vectors, and measuring the output quantum state in the computational basis. The resulting probability distribution is square-rooted and mapped back to the closest amino acid token magnitudes, enabling the reconstruction of new peptide sequences.

Encoding Scheme

QO-BRA operates on primary amino acid sequences by converting each peptide into a unique real amplitude vector. These vectors are then element-wise square-rooted to produce normalized state vectors, which serve as quantum inputs to the model. All 20 canonical amino acids are represented in the encoding. To differentiate between metal-binding and nonbinding residues, two categories are defined: `AA`

https://doi.org/10.1021/acs.jctc.5c01704
J. Chem. Theory Comput. 2026, 22, 2059–2073

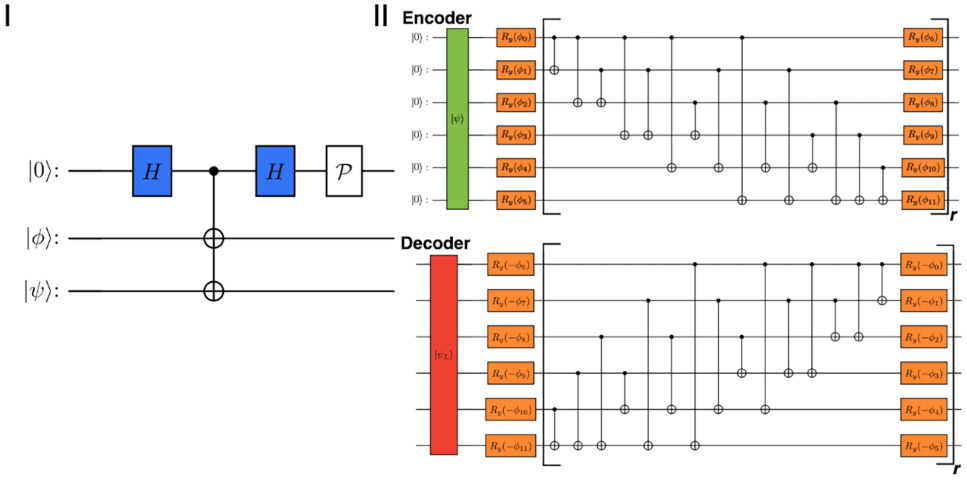


Figure 6. Panel I: Illustration of the SWAP test mechanism. The circuit ends with a probability measurement ($\mathcal{P}$) of the auxiliary qubit on top. When $\mathcal{P}(|0\rangle) = 1$, the states $|\phi\rangle$ and $|\psi\rangle$ are identical. Panel II: 6-qubit RealAmplitutes encoder and decoder ansätze are reported. The state vector input encoding layer (in green for the encoder and in red for the decoder). Rotation gates with trainable parameters are marked in orange. The repetition unit is highlighted with square brackets and the hyperparameter r.

refers to an amino acid that is not coordinated to a metal ion, while AA+ designates a metal-binding variant. Each AA and AA+ is assigned a unique numeric token.

Two special tokens are also introduced:

- : − Denotes chain breaks in multichain peptides.
- :X—Indicates the end of a sequence. Since the number of qubits (N_q) determines the dimensionality of the quantum state vector, sequences shorter than $2^{N_q} - 1$ residues are padded by appending :X, followed by repeated copies of the peptide. Sequences exceeding this length are truncated at the :X marker.

This encoding scheme establishes a consistent and reversible mapping from biological peptide sequences to fixed-dimensional quantum state vectors, enabling efficient quantum processing of peptides with diverse lengths and structural features.

Many classical machine learning models are invariant to the absolute values and ordering of input tokens.[71,72] However, the specific choice of token-to-value mapping significantly impacts the model performance of our quantum encoding. This is due to the sensitivity of the circuit to the input vector distribution. For instance, if two tokens with close numerical values—such as free Aspartic acid (D) and its ion-bound form (D+)—occur at similar frequencies in the training data, the encoder's intrinsic noise can lead to ambiguity between them. This results in an oversampling of the less frequent token due to value overlap in the latent space. To mitigate this, tokens are assigned numerical values that follow a bell-shaped distribution centered at zero, as shown in Figure S5. This ensures sufficient separation between tokens, especially for low-frequency ones. Additionally, because quantum measurements return probabilities—i.e., the squared amplitudes—any phase information (sign of the amplitude) is lost. To address this, all token values are assigned unique absolute magnitudes to preserve distinguishability.

Amplitude encoding is normalized, but peptide sequences may vary in length and total token value. To ensure a bijective and decodable representation, we prepend each vector with a fixed constant n. This scalar acts as an internal normalization reference, allowing for rescaling and accurate reconstruction of the original sequence. This format ensures compatibility with amplitude encoding while retaining biological interpretability. For example, the peptide sequence GC⋯LDAE is mapped as follows

$$\text{GC}\cdots\text{LDAE} \rightarrow [n \quad f(\text{G}) \quad f(\text{C})\cdots f(\text{L}) \quad f(\text{D}) \quad f(\text{A}) \quad f(\text{E})]^T$$

where $f()$ assigns a distinct real-valued amplitude to each input token, as defined by a given dictionary.

Encoder and Decoder

During training, QO-BRA maps the input data into a latent space representation by implementing unitary transformations.[24] The architecture of the encoder—i.e., the choice of ansätze—is critical, as it must balance expressiveness with trainability and hardware efficiency. Just as architectural choices define the learning capacity of a classical neural network, the quantum ansätze determine the representational power of the resulting QML model.

For qubit-based architectures, we employ the fully entangled RealAmplitudes (RA) ansätze (Figure 6II). In its Qiskit implementation,[45] it consists of successive layers of single-qubit R_y rotation gates interleaved with entangling layers characterized by all-to-all qubit connectivity. As an illustrative example employing a minimal 3-qubit circuit, Figure S6 contrasts the chosen entanglement configuration (I, red) with alternative topologies such as the ladder (II) and ring (III). The principal advantage of the selected scheme is that it permits maximal connectivity among the qubits, thereby enhancing the exchange and propagation of information throughout the circuit during the training process. All qubits are initialized in the computational basis state $|0\rangle$, and the exclusive use of R_y rotations constrains the evolution of the state vectors to the XZ-plane of the Bloch sphere. Consequently, all amplitudes remain real-valued, which in turn enables the use of real-valued loss functions.[73]

Figure 6II illustrates the minimal realization of the RA ansätze, specified by a repetition depth of $r = 1$. The representational capacity of this ansätze can be systematically enhanced by increasing r, which effectively appends additional RA layers and introduces further trainable parameters. Formally, the encoder ansätze U_E is defined as

$$U_E(\boldsymbol{\theta}) = \left(\prod_{i=1}^{r} U_y(\theta_E^i) \otimes U_{\text{ent}}\right) \otimes U_y(\theta_E^{r+1}) \tag{2}$$

where the ith layer is composed of a sequence of unitary entangling operations U_{ent} followed by a set of single-qubit rotational unitaries $U_y(\theta^i)$. The rotational unitary $U_y(\theta^i)$ consists of independent rotations about the y-axis on each qubit and is given by

$$U_y(\theta^i) = \bigotimes_{j=1}^{N_q} R_y(\theta^{i,j}) \tag{3}$$

with

$$R_y(\theta^{i,j}) = e^{-i\frac{\theta^{i,j}}{2}\hat{\sigma}_y} \tag{4}$$

https://doi.org/10.1021/acs.jctc.5c01704
J. Chem. Theory Comput. 2026, 22, 2059−2073

where $\hat{\sigma}_y$ denotes the Pauli-Y operator, which generates rotations about the y-axis. The entangling operator U_{ent} is implemented with all-to-all connectivity and is defined as

$$U_{\text{ent}} = \bigotimes_{j=1}^{N_q-1} \bigotimes_{k=j+1}^{N_q} CX(j, k) \tag{5}$$

where

$$CX(j, k) = I_j \otimes |0\rangle_k\langle 0|_k + X_j \otimes |1\rangle_k\langle 1|_k \tag{6}$$

denotes the controlled-X (CNOT) gate acting on qubits j (control) and k (target), thereby establishing quantum correlations between them. The decoder U_D is constructed structurally as the adjoint of the encoder circuit, i.e.

$$U_D(\boldsymbol{\phi}) = U_y(-\phi^{r+1}) \otimes \left(\prod_{i=1}^{r} U_{\text{ent}}^{\dagger} \otimes U_y(\phi^i)\right) \tag{7}$$

For the small-qubit regimes considered in this work ($N_q \leq 9$, corresponding to Hilbert-space dimension $2^{N_q} \leq 512$), forward-pass evaluations are accelerated by explicitly constructing the circuit unitary matrices and applying them via GPU-accelerated linear algebra. Specifically, after instantiating the parametrized encoder ansätze $U_E(\boldsymbol{\theta})$ (and the decoder U_D), we compute their corresponding unitary matrix representations and apply them to batches of amplitude-encoded input states using dense matrix−vector multiplications. With states stored as row vectors, this is implemented as $|\psi_L\rangle = U_E|\psi_{\text{in}}\rangle$ and $|\psi_{\text{rec}}\rangle = U_D|\psi_L\rangle$, which enables efficient batched evaluation in complex64 precision on the GPU. During phase 2 training, the encoder is kept fixed and its unitary matrix is cached in device memory, while only the decoder unitary is recomputed and updated at each iteration.

As aforementioned, the RealAmplitudes ansätze do not span the full $SU(2^{N_q})$ unitary manifold and, therefore, have reduced nominal expressivity; however, this constraint is imposed deliberately to promote optimization stability in data-limited regimes, preserve a structured and interpretable latent representation, and maintain circuit depths compatible with NISQ-scale evaluation. In contrast, fully expressive ansätze would substantially increase parameter counts, complicate the optimization landscape, and obscure latent-space structure.

The number of trainable parameters in the encoder scales as $N_q(r + 1)$, where N_q is the number of qubits. Unless otherwise specified, all results in this study are based on circuits with 6−9 qubits, enabling the representation of proteins with up to 63−511 amino acid residues. The decoder, illustrated in Figure 6II, is constructed as the complex conjugate (adjoint) of the encoder circuit. It receives the latent vector $|\psi_L\rangle$ as input and shares the same parameter values as the encoder. This architecture ensures efficient and consistent reconstruction in the quantum autoencoding pipeline.

Loss Functions

Losses in QO-BRA are constructed to (i) organize the learned quantum latent manifold into a tractable prior for generative sampling (Figure 1) and (ii) enforce faithful reconstruction during decoder training. We first state the classical m-MMD latent-regularization objective,[40,74] then introduce its SWAP-kernel quantum analogue that avoids explicit state tomography.[75] We finally describe the two-phase training protocol and the decoder-stage composite objective combining Hilbert-space fidelity and an ESM2-based biological regularizer.

MMD Loss. Classically, latent-prior alignment is commonly enforced using maximum mean discrepancy (MMD)[74] or its modified form (m-MMD).[40] For latent samples $\vec{x}_i$ (from the encoder) and reference vectors $\vec{y}_j$ drawn from a Gaussian prior, m-MMD is written as

$$\mathcal{L}(\vec{x}, \vec{y}) = 1 - \frac{1}{N^2}\sum_{i=0}^{N}\sum_{j=0}^{N} \mathcal{K}(\vec{x}_i, \vec{y}_j) \tag{8}$$

with a radial-basis kernel

$$\mathcal{K}(\vec{x}_i, \vec{y}_j) = \exp\left[-\frac{1}{2\sigma_{\text{kernel}}^2}\cdot\frac{1}{D}\sum_{d=0}^{D}(\vec{x}_{i_d} - \vec{y}_{j_d})^2\right] \tag{9}$$

where D is the latent dimensionality and σ_{kernel} is a bandwidth hyperparameter. In our amplitude-encoded setting, $\vec{y}_j$ are sampled and normalized to represent unit-norm latent states.

On quantum hardware, directly comparing state amplitudes would require tomography.[75] We therefore reformulate similarity in terms of state overlaps that admit a SWAP-test interpretation[76,77] (Figure 6I).

Starting from eq 8, we define a SWAP-kernel analogue by replacing the classical kernel similarity with a fidelity-derived dissimilarity

$$\mathcal{L}_{\text{SWAP}}(\vec{x}, \vec{y}) = \frac{1}{N^2}\sum_{i=0}^{N}\sum_{j=0}^{N} \mathcal{K}_{\text{SWAP}}(\vec{x}_i, \vec{y}_j) \tag{10}$$

$$\mathcal{K}_{\text{SWAP}}(\vec{x}_i, \vec{y}_j) = 1 - \left|\langle\psi_{\vec{x}_i}|\psi_{\vec{y}_j}\rangle\right|^2 \tag{11}$$

This yields a quantum analogue of the m-MMD objective used in kernel-elastic autoencoders,[40] while relying only on overlap estimates.

An explicit implementation of the SWAP test requires controlled-SWAP operations, resulting in increased circuit depth and a reliance on multiple two-qubit gates, which are among the dominant sources of error on NISQ-era devices. Consequently, practical execution on near-term hardware would likely require error-mitigation strategies or depth-reduction techniques to obtain reliable fidelity estimates. During training, no ancilla qubits or SWAP-test operations are implemented; instead, the SWAP test is employed as a mathematical analogue for reconstruction fidelity rather than as an explicitly executed circuit. The simulator directly evaluates pairwise state fidelities between latent vectors produced by the encoder and sampled Gaussian reference states, yielding the same overlap quantity that would be estimated via a SWAP-test measurement in a hardware setting.

Equation 11 can be interpreted in terms of the SWAP test, where the probability of measuring the ancilla qubit in the $|0\rangle$state encodes the squared overlap between two quantum states, $P(|0\rangle) = \frac{1}{2} + \frac{1}{2}\left|\langle\psi_{\vec{x}_i}|\psi_{\vec{y}_j}\rangle\right|^2$. $\mathcal{K}_{\text{SWAP}}$ therefore quantifies one minus this overlap, providing a fidelity-based measure of dissimilarity. We further note that SWAP-test−based fidelity estimation scales poorly with system size and circuit depth and would likely require alternative estimators (e.g., classical shadows or local overlap measurements) for practical deployment on larger NISQ-scale devices.

Optimization was performed using Qiskit's COBYLA implementation,[78] a derivative-free optimizer well suited to variational quantum circuits where the objective landscape can be nonsmooth and gradient estimation may be costly or noise-sensitive. Training used a full-batch size of 6,000 sequences (i.e., the entire training set) with a maximum of 500 epochs, and was terminated when the change in the objective fell below COBYLA's tolerance threshold or when the epoch limit was reached. Full-batch evaluation keeps the objective deterministic and helps amortize expensive circuit evaluations in the small-parameter regime; in contrast, stochastic mini-batching would introduce avoidable noise that can hinder convergence for gradient-free optimization. Consistent with prior quantum-autoencoder demonstrations,[79] no parameter-shift rule or backpropagation through the quantum circuit is used in the present work. All reported results are obtained from multiple random parameter initializations that converge to comparable solutions in noise-free simulation.

Reconstruction Fidelity Loss. After the convergence of the latent-space optimization, the encoder parameters are frozen, and a second training phase is initiated in which only the decoder parameters are optimized. In this phase, QO-BRA minimizes a composite loss consisting of a quantum-state reconstruction term and a protein-language-model regularization term. The quantum recon-

https://doi.org/10.1021/acs.jctc.5c01704
J. Chem. Theory Comput. 2026, 22, 2059−2073

struction loss is defined as a state-vector fidelity between the input quantum state and the output of the encoder−decoder circuit

$$\mathcal{L}_{\text{fid}} = 1 - |\langle\psi_{\text{in}}|U_D(\boldsymbol{\phi})U_E(\boldsymbol{\theta})|\psi_{\text{in}}\rangle|^2 \tag{12}$$

where $U_E(\boldsymbol{\theta})$ and $U_D(\boldsymbol{\phi})$ denote the independently parametrized encoder and decoder unitaries. This term enforces true Hilbert-space autoencoding, penalizing deviations between the reconstructed and original quantum states rather than relying on kernel-based similarity measures.

ESM Loss. To impose biological plausibility on decoded sequences, we introduce an ESM loss based on masked pseudolog-likelihoods computed using the pretrained protein language model ESM2_t6_8M_UR50D.[68] For each decoded sequence of length L, up to $K = \min(32, 0.15L)$ positions are randomly masked and scored under the ESM2 model, yielding a pseudolog-likelihood estimate of evolutionary consistency. To remove global scale dependence, the raw ESM loss is normalized by subtracting the corresponding baseline value computed once from the training set, so that zero corresponds to training-level plausibility. This produces a biologically grounded regularizer that rewards decoded sequences whose statistical properties resemble those of natural metalloproteins.

The decoder-stage objective is therefore

$$\mathcal{L}_{\text{dec}} = \lambda_{\text{fid}}\mathcal{L}_{\text{fid}} + \lambda_{\text{ESM}}(t)\mathcal{L}_{\text{ESM}} \tag{13}$$

where λ_{fid} controls the strength of quantum-state reconstruction and $\lambda_{\text{ESM}}(t)$ weights the contribution of the protein-language-model constraint. The ESM coupling is introduced using a stepwise schedule with an interval of 400 optimization steps: during the initial portion of decoder training, $\lambda_{\text{ESM}}(t) = 0$, allowing the decoder to first learn faithful quantum reconstruction, after which λ_{ESM} is switched to its target value to enforce biological plausibility. This staged coupling prevents premature collapse toward ESM-favored but poorly reconstructed solutions and ensures that QO-BRA first learns a coherent quantum autoencoder before incorporating evolutionary constraints.

ESM2 Evaluation

To assess whether the generated sequences recapitulate training-like residue-level statistics under an independent protein language model, we computed mean per-residue log-likelihoods using the ESM2 framework.[68] For each Zn^{2+} sequence ensemble (training set, QO-BRA, CNN-VAE, and random sampling), all sequences were first standardized to a fixed maximum length ($2^{N_q} = 512$) using the same padding and truncation scheme employed during QO-BRA training, thereby ensuring consistent tokenization and normalization across all methods. Each sequence was subsequently evaluated by ESM2_t30_150M_UR50D[68] to obtain a per-residue log-likelihood, and the resulting distributions were visualized as normalized histograms (Figure S4). The model was chosen for being the largest, most accurate within device memory limit, and this procedure interrogates higher-order sequence plausibility beyond simple token frequency statistics, as ESM2 assigns likelihoods based on evolutionary and structural regularities learned from large-scale protein sequence corpora.

Moreover, to quantitatively characterize discrepancies among these ESM2-derived distributions, we computed all pairwise two-sample Kolmogorov−Smirnov (KS) statistics[69] and one-dimensional Wasserstein distances[70] for every model combination (Table 6). The KS statistic captures the maximum absolute difference between the corresponding empirical cumulative distribution functions, whereas the Wasserstein distance quantifies the minimal "mass transport" required to transform one distribution into the other. Consequently, the Wasserstein distance provides a metric that is sensitive to both shifts in central tendency and changes in overall distributional dispersion.

Data Set

We used three curated data sets, publicly available via the QO-BRA GitHub repository,[80] each corresponding to protein complexes binding a specific divalent metal ion: Ca^{2+}, Mg^{2+}, or Zn^{2+}. The data sets include 13,279 proteins with Ca^{2+}-binding, 16,506 with Mg^{2+}-binding, and 19,474 proteins that bind Zn^{2+}. Structures were retrieved from the RCSB Protein Data Bank (PDB)[81] by filtering for entries annotated as containing the respective metal ion.

Because RCSB annotations often conflate structural cofactors with loosely associated ions,[82,83] we applied a postprocessing step to remove entries in which the metal ion was not covalently or coordinately bound to amino acid residues. All data sets were partitioned into training and testing sets using a 5:1 ratio.

CONCLUSION

We introduced the QO-BRA (Quantum Operator-Based Real Amplitude) autoencoder, a hybrid quantum−classical generative framework that leverages a variational quantum autoencoder to learn a latent representations of protein sequences and to enable controlled sampling of novel metalloprotein sequences. In QO-BRA, an encoder circuit maps amplitude-encoded inputs into the latent subspace, while a decoder realizes structurally as the adjoint operator enforces an explicit reconstructive symmetry that limits parameter growth and stabilizes training.

To make the latent space generative, we adopted a two-stage training strategy: (i) latent alignment using a distribution-matching objective (MMD) to encourage a tractable prior over latent variables, followed by (ii) decoder refinement using a composite objective that balances quantum-state reconstruction fidelity with a protein-language-model regularizer. The latter couples an ESM2-based plausibility term on a stepwise schedule, allowing the model to first learn faithful quantum reconstruction before enforcing sequence-level biological constraints. This staged coupling reduces collapse toward language-model-favored outputs that fail to reconstruct the underlying quantum representation.

Across curated Ca^{2+}-, Mg^{2+}-, and Zn^{2+}-binding data sets, we evaluated QO-BRA using complementary sequence- and structure-level probes. At the sequence level, we quantified distributional agreement under an independent ESM2 evaluation model and summarized pairwise discrepancies using two-sample KS statistics and Wasserstein distances. At the structure level, predicted and refined three-dimensional models were assessed using standard structural quality and biophysical descriptors (e.g., N-DOPE-based metrics, secondary-structure proportions, coordination environments), enabling a unified view of whether generated sequences preserve metalloprotein-like features beyond token-frequency matching.

Taken together, these findings establish QO-BRA as a concrete framework for integrating quantum autoencoding with biologically motivated regularization in protein sequence generation, while simultaneously elucidating the current limitations that impede scalability. In particular, fidelity estimation based on SWAP-style constructions is most appropriately regarded as a conceptual tool for small-scale systems and will likely require replacement by alternative estimators (e.g., local overlap-based metrics or methods inspired by classical shadows) as the number of qubits and the circuit depth increase. More broadly, future research directions include enhancing hardware realism (through noise-aware objective functions and measurement-efficient training protocols), extending conditional control beyond metal identity, and coupling sequence generation directly to downstream functional constraints (such as binding-site geometry and stability proxies). The ultimate aim is to progress from generating merely "plausible" sequences to generating sequences that are plausibly functional−the

https://doi.org/10.1021/acs.jctc.5c01704
J. Chem. Theory Comput. 2026, 22, 2059−2073

methodological analogue of advancing from compliance to convergence.

■ ASSOCIATED CONTENT

Data Availability Statement

The code and data are available at the QO-BRA Github.[80] For inquiries, please contact victor.batista@yale.edu.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.5c01704.

Experimental procedures and characterization data for all new compounds (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Victor S. Batista − *Department of Chemistry, Yale University, New Haven, Connecticut 06511, United States; Yale Quantum Institute, Yale University, New Haven, Connecticut 06511, United States;* orcid.org/0000-0002-3262-1237; Email: victor.batista@yale.edu

Authors

Yue Yu − *School of Engineering & Applied Sciences and Department of Chemistry, Yale University, New Haven, Connecticut 06511, United States; Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, Connecticut 06520, United States;* orcid.org/0000-0002-5509-2235

Francesco Calcagno − *Department of Industrial Chemistry "Toso Montanari" and Center for Chemical Catalysis−C3, University of Bologna, Bologna 40129, Italy;* orcid.org/0000-0002-0986-4425

Haote Li − *Department of Chemistry, Yale University, New Haven, Connecticut 06511, United States;* orcid.org/0000-0002-8146-5066

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.5c01704

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Y.Y. acknowledges Chuzhi Xu for providing a randomly generated protein sequence dataset used as a baseline comparison for ESM2 log-likelihood. V.S.B. acknowledges partial support from Boehringer Ingelheim; from the National Science Foundation Engines Development Award: Advancing Quantum Technologies (CT) under Award Number 2302908; and from the National Science Foundation Center for Quantum Dynamics on Modular Quantum Devices (CQD-MQD) under Award Number 2124511. We also acknowledge computational resources from the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility, located at Lawrence Berkeley National Laboratory. F.C. acknowledges the "Ing. Luciano Toso Montanari", Foundation for financially supporting his secondment at Yale University (USA) in the group of Professor Victor S. Batista.

■ REFERENCES

(1) Kuntz, I. D.; Meng, E. C.; Shoichet, B. K. Structure-based molecular design. *Acc. Chem. Res.* **1994**, *27*, 117−123.

(2) Freeze, J. G.; Kelly, H. R.; Batista, V. S. Search for catalysts by inverse design: artificial intelligence, mountain climbers, and alchemists. *Chem. Rev.* **2019**, *119*, 6595−6612.

(3) Gani, R.; Constantinou, L. Molecular structure based estimation of properties for process design. *Fluid Phase Equilib.* **1996**, *116*, 75−86.

(4) Ng, L. Y.; Chong, F. K.; Chemmangattuvalappil, N. G. Challenges and opportunities in computer-aided molecular design. *Comput. Chem. Eng.* **2015**, *81*, 115−129.

(5) Eslick, J. C.; Ye, Q.; Park, J.; Topp, E. M.; Spencer, P.; Camarda, K. V. A computational molecular design framework for crosslinked polymer networks. *Comput. Chem. Eng.* **2009**, *33*, 954−963.

(6) Pavurala, N.; Achenie, L. E. Identifying polymer structures for oral drug delivery-A molecular design approach. *Comput. Chem. Eng.* **2014**, *71*, 734−744.

(7) Contreras, M. L.; Rozas, R.; Valdivia, R. Exhaustive generation of organic isomers. 3. Acyclic, cyclic, and mixed compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 610−616.

(8) Davidson, S. Fast generation of an alkane-series dictionary ordered by side-chain complexity. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 147−156.

(9) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493−500.

(10) Shee, Y.; Li, H.; Zhang, P.; Nikolic, A. M.; Lu, W.; Kelly, H. R.; Manee, V.; Sreekumar, S.; Buono, F. G.; Song, J. J.; et al. Site-specific template generative approach for retrosynthetic planning. *Nat. Commun.* **2024**, *15*, No. 7818.

(11) Ramsundar, B. Molecular Machine Learning with DeepChem. Ph.D. thesis, Stanford University, 2018.

(12) De Cao, N.; Kipf, T. MolGAN: An Implicit Generative Model for Small Molecular Graphs. 2018, arXiv:1805.11973. arXiv.org e-Printarchive. https://arxiv.org/abs/1805.11973.

(13) Calcagno, F.; Serfilippi, L.; Franceschelli, G.; Garavelli, M.; Musolesi, M.; Rivalta, I. Quantum Chemistry Driven Molecular Inverse Design with Data-free Reinforcement Learning. 2025, arXiv:2503.12653. arXiv.org e-Printarchive. https://arxiv.org/abs/2503.12653.

(14) Dauparas, J.; Anishchenko, I.; Bennett, N.; Bai, H.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I.; Courbet, A.; de Haas, R. J.; Bethel, N.; et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **2022**, *378*, 49−56.

(15) Bruice, T. C.; Benkovic, S. J. Chemical basis for enzyme catalysis. *Biochemistry* **2000**, *39*, 6267−6274.

(16) Chen, Y.; Shanmugam, S. K.; Dalbey, R. E. The principles of protein targeting and transport across cell membranes. *Protein J.* **2019**, *38*, 236−248.

(17) Shin, Y. S.; Remacle, F.; Fan, R.; Hwang, K.; Wei, W.; Ahmad, H.; Levine, R.; Heath, J. R. Protein signaling networks from single cell fluctuations and information theory profiling. *Biophys. J.* **2011**, *100*, 2378−2386.

(18) Krebs, E. G. Protein phosphorylation and cellular regulation I. *Biosci. Rep.* **1993**, *13*, 127−142.

(19) Zhang, H.; Zhang, Z.; Guo, T.; Chen, G.; Liu, G.; Song, Q.; Li, G.; Xu, F.; Dong, X.; Yang, F.; et al. Annexin A protein family: Focusing on the occurrence, progression and treatment of cancer. *Front. Cell Dev. Biol.* **2023**, *11*, No. 1141331.

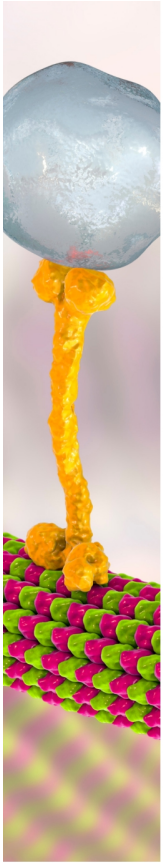
(20) Yang, Y.-H.; Wen, R.; Yang, N.; Zhang, T.-N.; Liu, C.-F. Roles of protein post-translational modifications in glucose and lipid metabolism: mechanisms and perspectives. *Mol. Med.* **2023**, *29*, No. 93.

(21) Dobson, C. M. The structural basis of protein folding and its links with human disease. *Philos. Trans. R. Soc. London* **2001**, *356*, 133−145.

(22) Listov, D.; Goverde, C. A.; Correia, B. E.; Fleishman, S. J. Opportunities and challenges in design and optimization of protein function. *Nat. Rev. Mol. Cell Biol.* **2024**, *25*, 639−653.
(23) Yu, Y.; Wang, R.; Teo, R. D. Machine learning approaches for metalloproteins. *Molecules* **2022**, *27*, No. 1277.
(24) Smaldone, A. M.; Shee, Y.; Kyro, G. W.; Xu, C.; Vu, N. P.; Dutta, R.; Farag, M. H.; Galda, A.; Kumar, S.; Kyoseva, E.; Batista, V. S. Quantum machine learning in drug discovery: Applications in academia and pharmaceutical industries. *Chem. Rev.* **2025**, *125*, 5436−5460.
(25) Greener, J. G.; Moffat, L.; Jones, D. T. Design of metalloproteins and novel protein folds using variational autoencoders. *Sci. Rep.* **2018**, *8*, No. 16189.
(26) Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; et al. De novo design of protein structure and function with RFdiffusion. *Nature* **2023**, *620*, 1089−1100.
(27) 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.
(28) 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.
(29) Akrom, M.; Rustad, S.; Dipojono, H. K. Variational quantum circuit-based quantum machine learning approach for predicting corrosion inhibition efficiency of pyridine-quinoline compounds. *Mater. Today Quantum* **2024**, *2*, No. 100007.
(30) 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.
(31) Thomas, A. M.; Chen, Y.-C.; Valencia, H. O.; Jose, S. T.; Wu, R. QCA-MolGAN: Quantum Circuit Associative Molecular GAN with Multi-Agent Reinforcement Learning. 2025, arXiv:2509.05051. arXiv.org e-Printarchive. https://arxiv.org/abs/2509.05051.
(32) Czischek, S. *Neural-Network Simulation of Strongly Correlated Quantum Systems*; Springer Nature, 2020.
(33) Massoli, F. V.; Vadicamo, L.; Amato, G.; Falchi, F. A leap among entanglement and neural networks: A quantum survey. 2021, arXiv:2107.03313. arXiv.org e-Printarchive. https://arxiv.org/abs/2107.03313.
(34) Caro, M. C.; Huang, H.-Y.; Cerezo, M.; Sharma, K.; Sornborger, A.; Cincio, L.; Coles, P. J. Generalization in quantum machine learning from few training data. *Nat. Commun.* **2022**, *13*, No. 4919.
(35) Jia, Z.-A.; Yi, B.; Zhai, R.; Wu, Y.-C.; Guo, G.-C.; Guo, G.-P. Quantum neural network states: A brief review of methods and applications. *Adv. Quantum Technol.* **2019**, *2*, No. 1800077.
(36) Khoshaman, A.; Vinci, W.; Denis, B.; Andriyash, E.; Sadeghi, H.; Amin, M. H. Quantum variational autoencoder. *Quantum Sci. Technol.* **2019**, *4*, No. 014001.
(37) Wang, G.; Warrell, J.; Emani, P. S.; Gerstein, M. Quantum variational autoencoder utilizing regularized mixed-state latent representations. *Phys. Rev. A* **2025**, *111*, No. 042416.
(38) Wu, J.; Fu, H.; Zhu, M.; Zhang, H.; Xie, W.; Li, X.-Y. Quantum circuit autoencoder. *Phys. Rev. A* **2024**, *109*, No. 032623.
(39) Rocchetto, A.; Grant, E.; Strelchuk, S.; Carleo, G.; Severini, S. Learning hard quantum distributions with variational autoencoders. *npj Quantum Inf.* **2018**, *4*, No. 28.
(40) Li, H.; Shee, Y.; Allen, B.; Maschietto, F.; Morgunov, A.; Batista, V. Kernel-elastic autoencoder for molecular design. *PNAS Nexus* **2024**, *3*, No. 168.
(41) Shee, Y.; Morgunov, A.; Li, H.; Batista, V. S. DirectMultiStep: Direct Route Generation for Multistep Retrosynthesis. *J. Chem. Inf. Model.* **2025**, *65*, 3903−3914.
(42) Mansimov, E.; Mahmood, O.; Kang, S.; Cho, K. Molecular geometry prediction using a deep generative graph neural network. *Sci. Rep.* **2019**, *9*, No. 20381.
(43) Dutta, R.; Cabral, D. G.; Lyu, N.; Vu, N. P.; Wang, Y.; Allen, B.; Dan, X.; Cortiñas, R. G.; Khazaei, P.; Smart, S. E.; et al. Simulating Chemistry on Bosonic Quantum Devices. *J. Chem. Theory Comput.* **2024**, *20*, 6426−6441.
(44) Romero, J.; Olson, J. P.; Aspuru-Guzik, A. Quantum autoencoders for efficient compression of quantum data. *Quantum Sci. Technol.* **2017**, *2*, No. 045001.
(45) 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.; Johnson, B. R.; Gambetta, J. M. Quantum computing with Qiskit. 2024, arXiv:2405.08810. arXiv.org e-Printarchive. https://arxiv.org/abs/2405.08810.
(46) Basha, S. S.; Dubey, S. R.; Pulabaigari, V.; Mukherjee, S. Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing* **2020**, *378*, 112−119.
(47) Huang, H.-Y.; Broughton, M.; Mohseni, M.; Babbush, R.; Boixo, S.; Neven, H.; McClean, J. R. Power of data in quantum machine learning. *Nat. Commun.* **2021**, *12*, No. 2631.
(48) Team, C. D.; Boitreaud, J.; Dent, J.; McPartlon, M.; Meier, J.; Reis, V.; Rogozhonikov, A.; Wu, K.Chai-1: Decoding the molecular interactions of life *BioRxiv* 2024. 2024-10.
(49) Eastman, P.; Galvelis, R.; Peláez, R. P.; Abreu, C. R.; Farr, S. E.; Gallicchio, E.; Gorenko, A.; Henry, M. M.; Hu, F.; Huang, J.; et al. OpenMM 8: molecular dynamics simulation with machine learning potentials. *J. Phys. Chem. B* **2024**, *128*, 109−116.
(50) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(51) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(52) Zhang, Z.; Liu, X.; Yan, K.; Tuckerman, M. E.; Liu, J. Unified efficient thermostat scheme for the canonical ensemble with holonomic or isokinetic constraints via molecular dynamics. *J. Phys. Chem. A* **2019**, *123*, 6056−6079.
(53) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528−1532.
(54) Rosato, A.; Valasatava, Y.; Andreini, C. Minimal functional sites in metalloproteins and their usage in structural bioinformatics. *Int. J. Mol. Sci.* **2016**, *17*, No. 671.
(55) Chalkley, M. J.; Mann, S. I.; DeGrado, W. F. De novo metalloprotein design. *Nat. Rev. Chem.* **2022**, *6*, 31−50.
(56) Cook, W. J.; Walter, L. J.; Walter, M. R. Drug binding by calmodulin: crystal structure of a calmodulin-trifluoperazine complex. *Biochemistry* **1994**, *33*, 15259−15265.
(57) Yamagami, R.; Bingaman, J. L.; Frankel, E. A.; Bevilacqua, P. C. Cellular conditions of weakly chelated magnesium ions strongly promote RNA stability and catalysis. *Nat. Commun.* **2018**, *9*, No. 2149.
(58) Lippard, S. *Principles of Bioinorganic Chemistry*; University Science Book, 1994; Vol. 2.
(59) Enamullah, M.; Quddus, M. A.; Halim, M. A.; Islam, M. K.; Vasylyeva, V.; Janiak, C. Switching from 4+ 1 to 4+ 2 zinc coordination number through the methyl group position on the pyridyl ligand in the geometric isomers bis [N-2-(4/6-methyl-pyridyl) salicylaldiminato-κ2N, O] zinc (II). *Inorg. Chim. Acta* **2015**, *427*, 103−111.
(60) Hekkelman, M. L.; Álvarez Salmoral, D.; Perrakis, A.; Joosten, R. P.DSSP 4: FAIR annotation of protein secondary structure *bioRxiv* 2025, 2025-04.
(61) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; de Hoon, M. J. L. Biopython: freely available Python tools for

https://doi.org/10.1021/acs.jctc.5c01704
J. Chem. Theory Comput. 2026, 22, 2059−2073

computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422−1423.
(62) Chakravarty, D.; Schafer, J. W.; Chen, E. A.; Thole, J. F.; Ronish, L. A.; Lee, M.; Porter, L. L. AlphaFold predictions of fold-switched conformations are driven by structure memorization. *Nat. Commun.* **2024**, *15*, No. 7296.
(63) McDonald, E. F.; Jones, T.; Plate, L.; Meiler, J.; Gulsevin, A. Benchmarking AlphaFold2 on peptide structure prediction. *Structure* **2023**, *31*, 111−119.
(64) Shen, M.-y.; Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **2006**, *15*, 2507−2524.
(65) Eramian, D.; Eswar, N.; Shen, M.-Y.; Sali, A. How well can the accuracy of comparative protein structure models be predicted? *Protein Sci.* **2008**, *17*, 1881−1893.
(66) Sawarkar, K. *Deep Learning with PyTorch Lightning*; Packt Publishing, 2022.
(67) Vadar, P. S.; Moharekar, T. T.; Pol, U. R. A Comprehensive Comparison of Deep Learning Libraries: TensorFlow, PyTorch, FastAI, Keras, and PyTorch Lightning. *Int. J. Comput. Sci. Eng. Tech.* **2024**, *8*, No. 6.
(68) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123−1130.
(69) Lilliefors, H. W. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *J. Am. Stat. Assoc.* **1967**, *62*, 399−402.
(70) Panaretos, V. M.; Zemel, Y. Statistical aspects of Wasserstein distances. *Annu. Rev. Stat. Appl.* **2019**, *6*, 405−431.
(71) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.
(72) Quinlan, J. *C4.5: Programs for Machine Learning; Ebrary Online*, Morgan Kaufmann 2014.
(73) He, G. P. Training Quantum Machine Learning Model on Cloud without Uploading the Data. 2024, arXiv:2409.04602. arXiv.org e-Printarchive. https://arxiv.org/abs/2409.04602.
(74) Lin, W.-W.; Mak, M.-W.; Li, L.; Chien, J.-T. Reducing domain mismatch by maximum mean discrepancy based autoencoders. *Odyssey* **2018**, 162−167.
(75) O'Donnell, R.; Wright, J. In *Efficient quantum tomography*, Proceedings of the forty-eighth annual ACM symposium on Theory of Computing; ACM Digital Library, 2016; pp 899−912.
(76) Barenco, A.; Berthiaume, A.; Deutsch, D.; Ekert, A.; Jozsa, R.; Macchiavello, C. Stabilization of quantum computations by symmetrization. *SIAM J. Comput.* **1997**, *26*, 1541−1557.
(77) Kang, M.-S.; Heo, J.; Choi, S.-G.; Moon, S.; Han, S.-W. Implementation of SWAP test for two unknown states in photons via cross-Kerr nonlinearities under decoherence effect. *Sci. Rep.* **2019**, *9*, No. 6167.
(78) Pellow-Jarman, A.; Sinayskiy, I.; Pillay, A.; Petruccione, F. A comparison of various classical optimizers for a variational quantum linear solver. *Quantum Inf. Process.* **2021**, *20*, No. 202.
(79) Sahin, E.; Altamura, E.; Wallis, O.; Schiavello, F. Quantum Autoencoder 2024; https://github.com/qiskit-community/qiskit-machine-learning/blob/stable/0.8/docs/tutorials/12_quantum_autoencoder.ipynb.
(80) Yu, Y. *QO-BR: Quantum Operator-Based Real Amplitude Autoencoder*, 2025; https://github.com/SamuelYueYu/QO-BRA_1.0.git.
(81) Rose, P. W.; Bi, C.; Bluhm, W. F.; Christie, C. H.; Dimitropoulos, D.; Dutta, S.; Green, R. K.; Goodsell, D. S.; Prlić, A.; Quesada, M.; et al. The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.* **2012**, *41*, D475−D482.
(82) Harding, M. M. Metal-ligand geometry relevant to proteins and in proteins: sodium and potassium. *Acta Crystallogr., Sect. D:Biol. Crystallogr.* **2002**, *58*, 872−874.
(83) Zheng, H.; Chruszcz, M.; Lasota, P.; Lebioda, L.; Minor, W. Data mining of metal ion environments present in protein structures. *J. Inorg. Biochem.* **2008**, *102*, 1765−1776.



CAS BIOFINDER DISCOVERY PLATFORM™
BRIDGE BIOLOGY AND CHEMISTRY FOR FASTER ANSWERS
Analyze target relationships, compound effects, and disease pathways
Explore the platform
CAS
A division of the American Chemical Society

https://doi.org/10.1021/acs.jctc.5c01704
J. Chem. Theory Comput. 2026, 22, 2059−2073