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

Yue Yu [†,‡,¶] Francesco Calcagno [§,‖] Haote Li [¶] and Victor S. Batista [*,¶,⊥]

†*School of Engineering & Applied Science, Yale University, New Haven, CT 06511, USA*

‡*Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, CT 06520, USA*

¶*Department of Chemistry, Yale University, New Haven, CT 06511, USA*

§*Department of Industrial Chemistry "Toso Montanari", University of Bologna, Via Piero Gobetti 85, 40129 Bologna, Italy*

‖*Center for Chemical Catalysis – C3, University of Bologna, Via Piero Gobetti 85, 40129 Bologna, Italy*

⊥*Yale Quantum Institute, Yale University, New Haven, CT 06511, USA*

E-mail: victor.batista@yale.edu

Software and Hardware

All computations were executed using Python 3.11.5. The behavioral emulation of a quantum device via classical computation was performed using the Qiskit (version 1.2.0) quantum simulation package. Quantum state measurements were performed using the StateVector simulator.[1] Circuit parameter optimization was achieved using the COBYLA optimizer, as

implemented within Qiskit. Tensor calculations were facilitated by the PyTorch version 2.2.0 + cu121 package.[2] A comprehensive list of libraries used in this study is provided in this link. All simulations delineated in this study were performed utilizing 16 processors in the following hardware configuration: AMD Perlmutter EPYC CPUs equipped with 512 GB of RAM and NVIDIA A100 Tensor Core GPUs, featuring 40 GB of HBM2.

Selectivity and Specificity

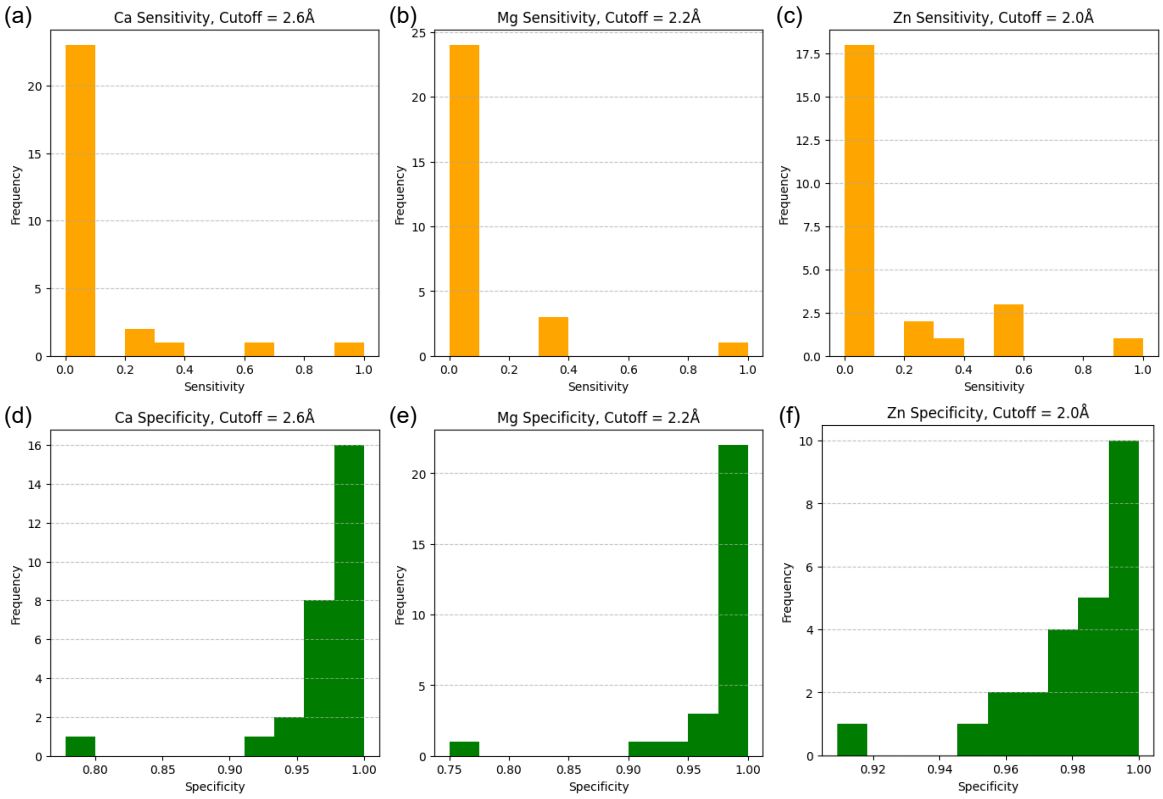

Figure S1: Selectivity and specificity among the generated sets of residues predicted to coordinate each metal ion type were evaluated. Overall, the model exhibits high specificity, with the majority of values falling within the 0.9–1.0 range for all three ions, indicating a low rate of false-positive predictions. In contrast, sensitivity displays substantially greater variability and frequently approaches zero, suggesting that many true coordinating residues are not identified. Nevertheless, instances of 100% sensitivity are observed, demonstrating that, under favorable structural conditions, the model can achieve near-perfect detection of coordinating residues.

Coordination Numbers

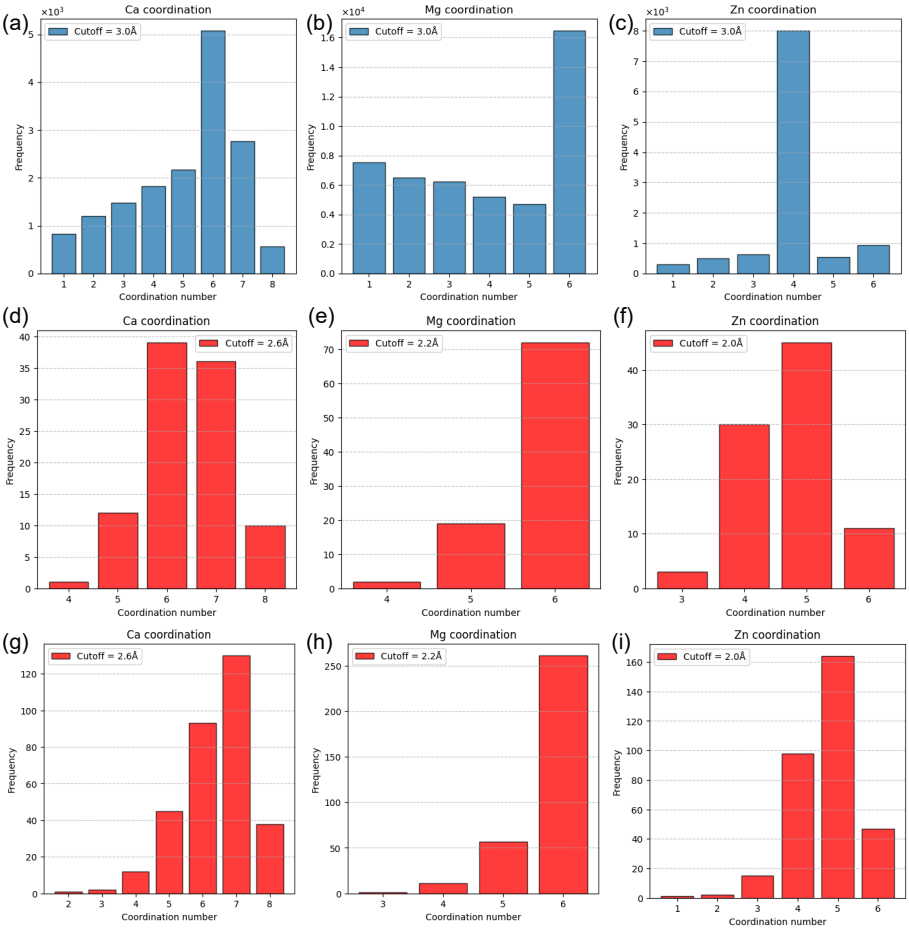

Figure S2: Comparison of the metal coordination number frequency for Ca^{2+}-, Mg^{2+}-, and Zn^{2+}-proteins in the real case (a-c) against generated structures containing 1 (d-f) and 4 (g-i) ions. The cutoffs are chosen so that the majority of coordination numbers match the theoretical values.

Normalized DOPE Score

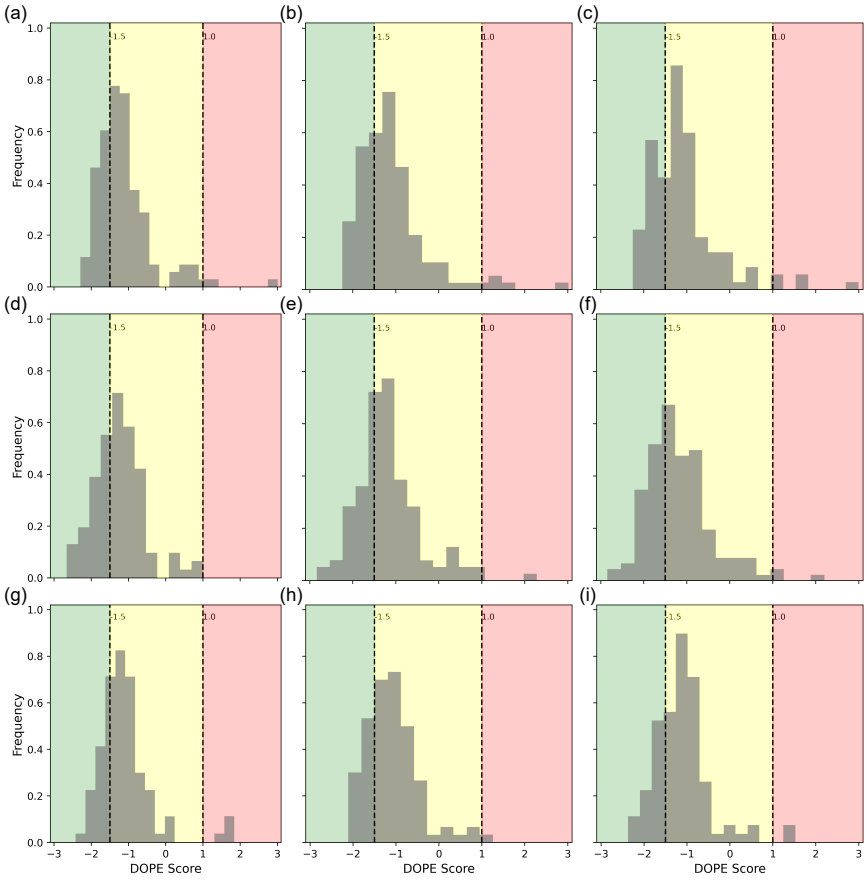

Figure S3: Distributions of N-DOPE scores for the generated structures of Ca^{2+}- (a–c), Mg^{2+}- (d–f), and Zn^{2+}-binding proteins (g–i) are shown, with the number of ions in the models increasing from one on the left to three on the right. Following the criteria of Eramian et al.,[3] models with N-DOPE scores below -1.5 are classified as near-native, whereas those with scores above 1.0 are considered inaccurate. Accordingly, the histograms are partitioned into green (near-native), red (inaccurate), and yellow (ambiguous) regions. Under this classification, only a small fraction of the generated structures fall into the inaccurate category, with the remaining models approximately evenly divided between ambiguous and near-native quality.

ESM Analysis

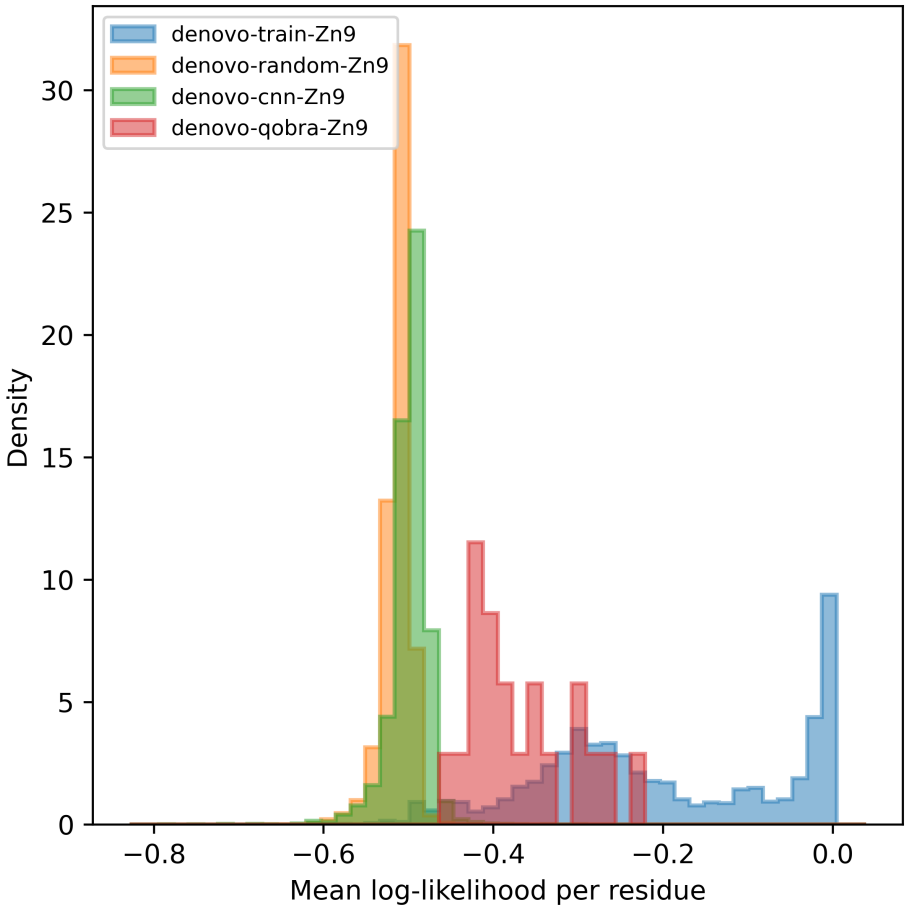

Figure S4: Comparison of mean per-residue log-likelihood distributions for the training set (blue) and sequences generated via random sampling (orange), a classical CNN-based VAE (green), and QO-BRA (red). Log-likelihoods were computed using ESM2_t30_150M_UR50D[4] for all sequences; for each sequence, per-position log-likelihoods were averaged to obtain a single summary value.

Encoding Scheme

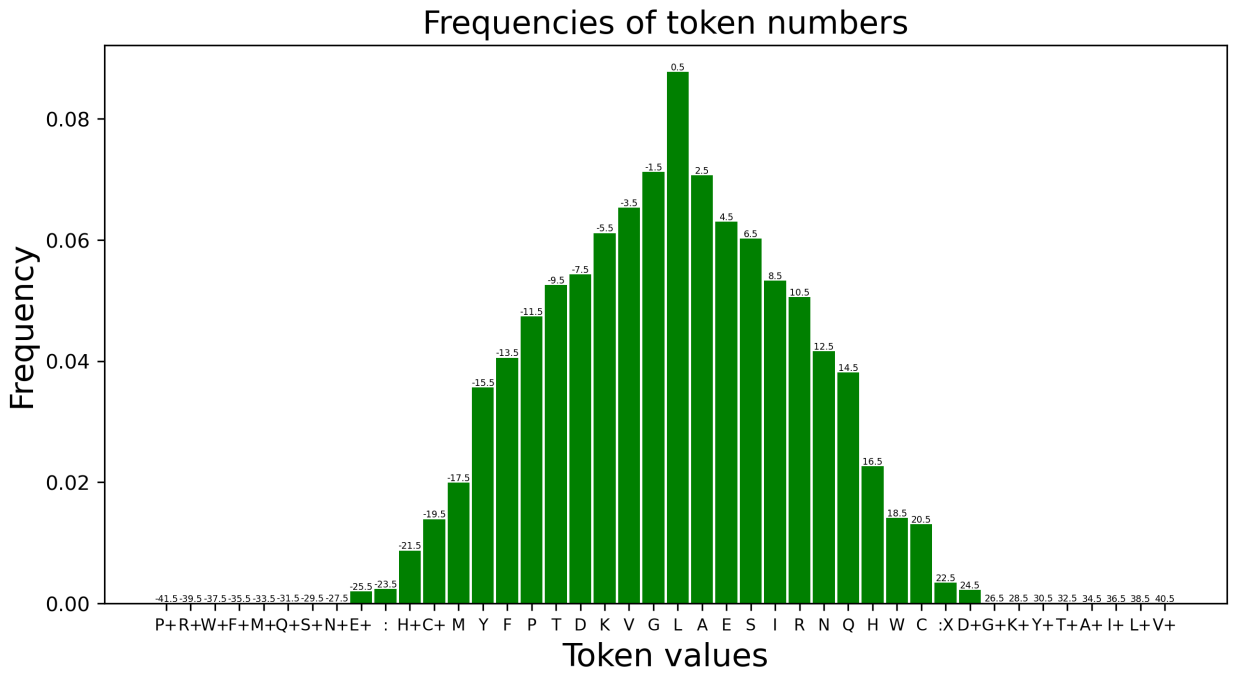

Figure S5: Assignment of token values to amino acid sequence characters, illustrated here using the Zn^{2+}–protein dataset as an example, is conducted as follows. The most frequent character is mapped to the token with the smallest absolute value, and subsequent characters are assigned token values of increasing absolute magnitude as their frequencies decrease, with signs alternating between positive and negative. The smallest absolute value is set to 0.5, and subsequent values increase in steps of 1.0 in absolute terms, ensuring that no two characters share the same absolute token value. This construction explicitly accounts for the loss of sign information in the measurement of quantum state probabilities on real quantum hardware. Consequently, the encoding scheme obviates the need for quantum state tomography to recover phase information.

Entanglement Scheme

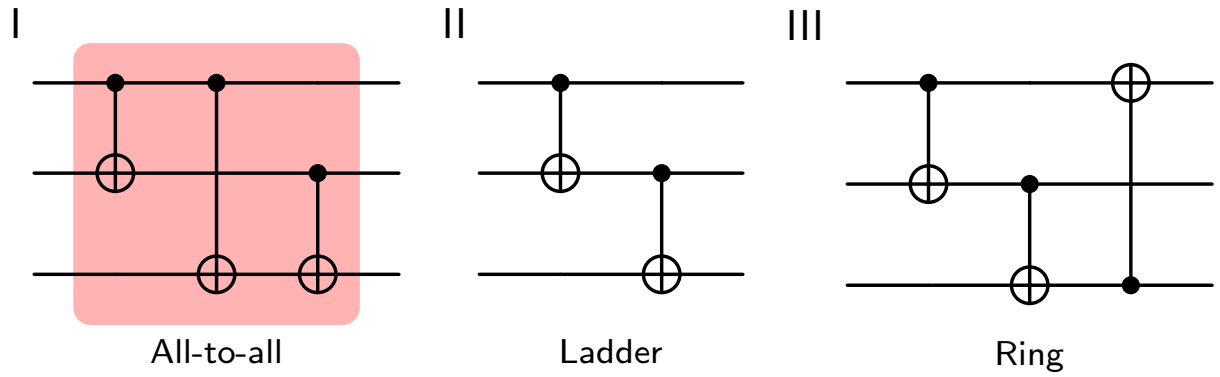

Figure S6: Demonstration of all-to-all qubit entanglement (I, red) used for QO-BRA illustrated with three qubits and compared against alternative ladder (II) and ring (III) entanglements. All-to-all allows for maximum exchange of information among the qubits, facilitating the learning process.

References

(1) 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.

(2) Imambi, S.; Prakash, K. B.; Kanagachidambaresan, G. PyTorch. *Programming with TensorFlow: solution for edge computing applications* **2021**, 87–104.

(3) Eramian, D.; Eswar, N.; Shen, M.-Y.; Sali, A. How well can the accuracy of comparative protein structure models be predicted? *Protein Science* **2008**, *17*, 1881–1893.

(4) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; others Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123–1130.