Protein-protein Docking Using a Tensor Train Black-box Optimization Method

Life Sciences

Published by Dmitry Morozov, Artem Melnikov, Vishal Shete, Michael Perelshtein

2023

terraquantum.swiss

Protein-protein docking using a tensor train black-box optimization method

Dmitry Morozov,* Artem Melnikov, Vishal Shete, and Michael Perelshtein†
Terra Quantum AG, Kornhausstrasse 25, 9000 St. Gallen, Switzerland

Black-box optimization methods play an important role in many fields of computational simulation. In particular, such methods are often used in the design and modelling of biological systems, including proteins and their complexes with various ligands. This work is mainly focused on the protein-protein docking that plays a key role in modern drug-design workflows. We develop a black-box approach for such docking problems using a novel technique based on the tensor-train decomposition of high-dimensional interaction functions. Our method shows an advantage in terms of the discovered global minima and has a high potential for further implementation on a wide range of devices, including graphical processing units and quantum processing units.

arXiv:2302.03410v1 [q-bio.BM] 7 Feb 2023

I. INTRODUCTION

Protein-protein interactions play a crucial role in organisms [1]. Understanding how and why individual molecules could form complexes of a very distinct structure is challenging. There are experimental methods to determine the structures of complexes, like crystallography and CryoEM techniques, which are typically deposited into the Protein Data Bank (PDB) [2]. However, they can cover only a very tiny fraction of all biological complexes. To overcome these limitations, computational methods for predicting and building protein complexes are becoming more popular, as a complementary addition to the experimental techniques.

Molecular docking is one of the major methods in studying protein-protein interactions at the atomistic scale [3, 4]. Essentially, the method aims to sample the geometrical space of two or more proteins and search for the best poses of one protein (the ligand) with respect to the other protein (the receptor) using scoring functions. The main challenges are: (i) to design the scoring function to give an acceptable precision; and (ii) to efficiently sample the conformational space to find the best possible structures.

Today's most widely used sampling methods are based on the Fast Fourier Transform (FFT). They are capable of generating geometrically complementary rigid-body docking poses [5]. However, they are typically limited in the use of very specific scoring functions and thus, in many cases require an *a posteriori* recalculation of the scores for the obtained structures with additional external software [6]. On the other hand, there are existing black-box approaches for searching minimas in high-dimensional spaces. As an example, swarm intelligence algorithms, such as Particle Swarm Optimization (PSO) [7] and Glowworm Swarm Optimization (GSO) methods [8], have been successfully applied to the protein docking problem using a large variety of scoring functions [9].

In this work, we are aiming to apply another class of black-box global optimization algorithms based on the tensor train (TT) decomposition of the high-dimensional scoring (or energy) function [10] that were developed over the last years [11]. The expected advantages of such methods is their agnosticism to the initial conditions and high-efficiency in sampling even very heterogeneous coordinate spaces that have a large number of local minima in their structure. These properties make such methods promising candidates for exploring interactions in biological complexes.

Another advantage of the tensor train methods developed by Terra Quantum is their capability for **efficient implementation on both CPUs and GPUs**. As can be seen from the results in this paper, usage of these algorithms already produces highly promising results. Moreover, the tensor train algorithms possess a structure similar to quantum circuits [12, 13], which paves the way towards their future implementation on Quantum Processing Units (QPUs) once such QPUs are of sufficient scale and fidelity [14]. Thus, we are able to harness some advantages of our *Quantum Software* approach [15], which will grow once quantum hardware matures and numerical methods of quantum information processing via tensor networks advances.

II. METHODS

Three different algorithms for solving docking problem were employed in the study: the glowworm swarm optimization (GSO) algorithm as implemented in the LightDock [16] package; the tensor train based minima search

* dimo@terraquantum.swiss
† mpe@terraquantum.swiss

implemented in the ttpy library (TTPY) [17]; and our own tensor train based black-box optimization method packaged in a Python library (TTOPT), which will be described in the following section.

A. Tensor Train assisted docking

Here we will describe our workflow for performing docking simulations using the optimization algorithm based on the Tensor Train (TT) decomposition [10, 11, 18]. The overall scheme is shown in Fig. 1. Initially, as in the other docking protocols, the preparation and cleanup of the ligand and receptor structures should be performed. At that stage, we also decide on the target (scoring) function used to evaluate the docking poses of the ligand molecule with respect to the receptor. This stage could also involve fixing the missing residues and side-chains, adding hydrogenes, building molecular topologies and partially relaxing the structures.

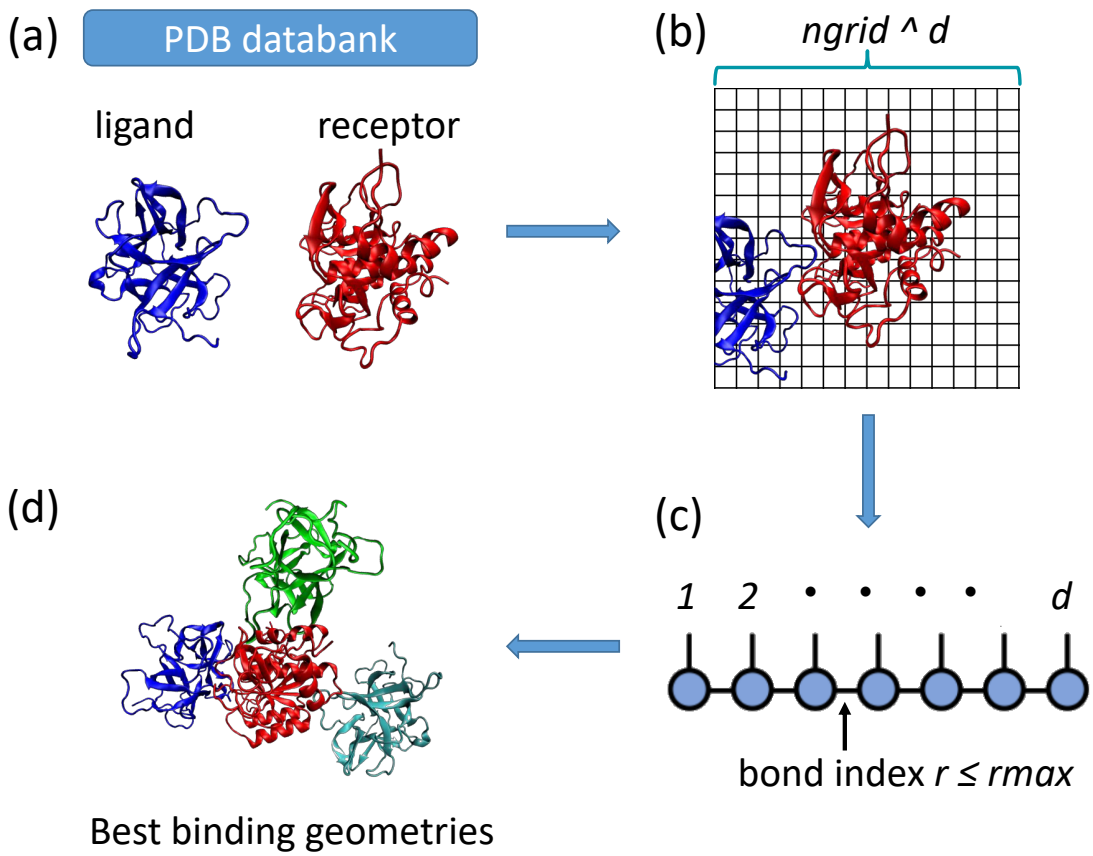

FIG. 1. Workflow for performing docking simulations with the TTOPT approach. (a) - structure preparation, (b) - grid and target function definition, (c) - decomposition of the target function into a tensor train and search for optima on a grid, (d) - clusterization of the obtained minima geometries.

For the next stage, one would need to define a grid spanned around the receptor Fig. 1b in which optimization of the target function with respect to the ligand position will be performed. It could be both minimization of the energy or maximization of the scoring. In general, our TT algorithm requires discretization of the search on a uniform grid. We define d as the number of variables (dimensions) and $ngrid$ as the grid size in one dimension. In the current work, we use rigid-body models of the receptor and ligand proteins meaning that the total search space has 7 coordinates in total: 3 coordinates for the ligand's center of masses and 4 coordinates for its rotation quaternion. It should be noted that while here we are focused mainly on rigid-body protein-protein docking, the same approach is also applicable to other systems such as small molecules and nucleic acids, where the flexibility of these systems could be straightforwardly treated as additional dimensions in the TT decomposition.

The TTOPT algorithm would be used to search for the possible optima of the target function in the defined grid space. Unlike brute-force grid search, this algorithm does not estimate the cost function at all points of the grid but dynamically provides the next set of evaluating points in the search space based on the knowledge accumulated during all previous evaluations. The TTOPT algorithm applies the cross-approximation technique to the tensors [19], which in turn is based on the `MaxVol` routine [20]. Essentially, it performs a decomposition of the multi-dimensional target function into a compressed form as the product of small tensors connected into a chain (Fig. 1c) and expanded on a grid.

In addition, the iterative power algorithm by Soley *et.al.*[21] is used for the global minima search within the TT representation. The total number of the target function evaluations required by our approach to find global minima could be approximated with $O(nswp * d * ngrid * rmax^2)$, where d is the number of dimensions, $ngrid$ is the grid-size, $rmax$ is the maximum rank of the tensors with which the algorithm tries to approximate the target function via tensor train decomposition and $nswp$ is the number of iterations which the algorithm performs to improve the decomposition. Overall, our TTOPT approach scales much better with the system size than the $ngrid^d$ brute force grid search and shows much better parallelization capabilities than other black-box methods, such as Bayesian optimization or genetic algorithms. We implement the TTOPT core functional using C++ and package the whole framework in a Python library.

Finally, after the global optima search, we perform an analysis of the obtained receptor-ligand complex structures in terms of the target function values and clusterize them according to the root mean square deviation (RMSD) of the ligand Cα protein backbone atoms (Fig. 1d). Obtained in such a way, the best complexes from all found clusters could be further used to perform a comparison with the experimental data or for conducting simulations such as molecular dynamics, binding free-energy evaluations or direct mutagenesis using more elaborate approaches.

B. Systems for docking

To perform docking simulations, we have used 10 systems from the CAPRI scoring benchmark [22]. The Protein Data Bank (PDB) structures along with the chain identifiers, which have been used as the receptor and ligand parts of the respecting complexes to perform the simulations, are summarized in Table I. The fast implementation of the DFIRE energy function [23] provided in the LightDock software package was used as a target function for performing global minima search.

TABLE I. Systems considered in this work. Chain - ID taken from the original PDB, Length - number of the residues in the Chain, Weight - molecular weight of the Chain in kDa

PDBID	Receptor			Ligand		
	Chain	Length	Weight	Chain	Length	Weight
3BX1	A	269	26.7	C	181	19.9
3U43	B	132	15.3	A	92	10.0
3R2X	AB	492	56.8	C	82	10.6
3Q87	B	164	19.2	A	122	14.2
2WPT	B	113	15.2	A	82	9.9
3E8L	A	223	23.3	C	176	20.2
3FM8	C	364	45.6	B	95	14.0
2W83	A	161	19.0	CD	129	17.7
2REX	B	171	21.9	A	107	13.3
2VDU	D	376	51.3	F	201	29.7

In the GSO optimization, 128 swarms have been simulated. Each swarm consisted of the 600 glowworms randomly placed at the beginning of the simulation. 200 algorithm iterations have been performed for each swarm. In addition, the best glowworm in each swarm has been minimized for 100 steps using the Powell minimization method [24] at each GSO iteration. That resulted in ~ 140000 (i.e., $600 * 200 + 200 * 100$) total evaluations of the target function.

The tensor train based methods used the same parameters for the generation of the target function approximation with: the maximum rank of the tensor-train expansion $rmax = 3$; the number of grid points $ngrid = 64$ along each of the 7 dimensions (3 for the center of masses translation and 4 for the rotation quaternion); and the total number of iterations $nswp = 10$. The parameters were chosen such that the methods resulted in ~ 120000 target function evaluations. To be consistent with the GSO algorithm, we performed the TT-based minima search 128 times as well, starting with different random number seeds. In addition, for the TTPY method, we performed local energy minimization of the final structures for 100 steps using the same Powell approach. In TTOPT, we performed additional local minimizations for 20 steps for the best found points at the end of each iteration.

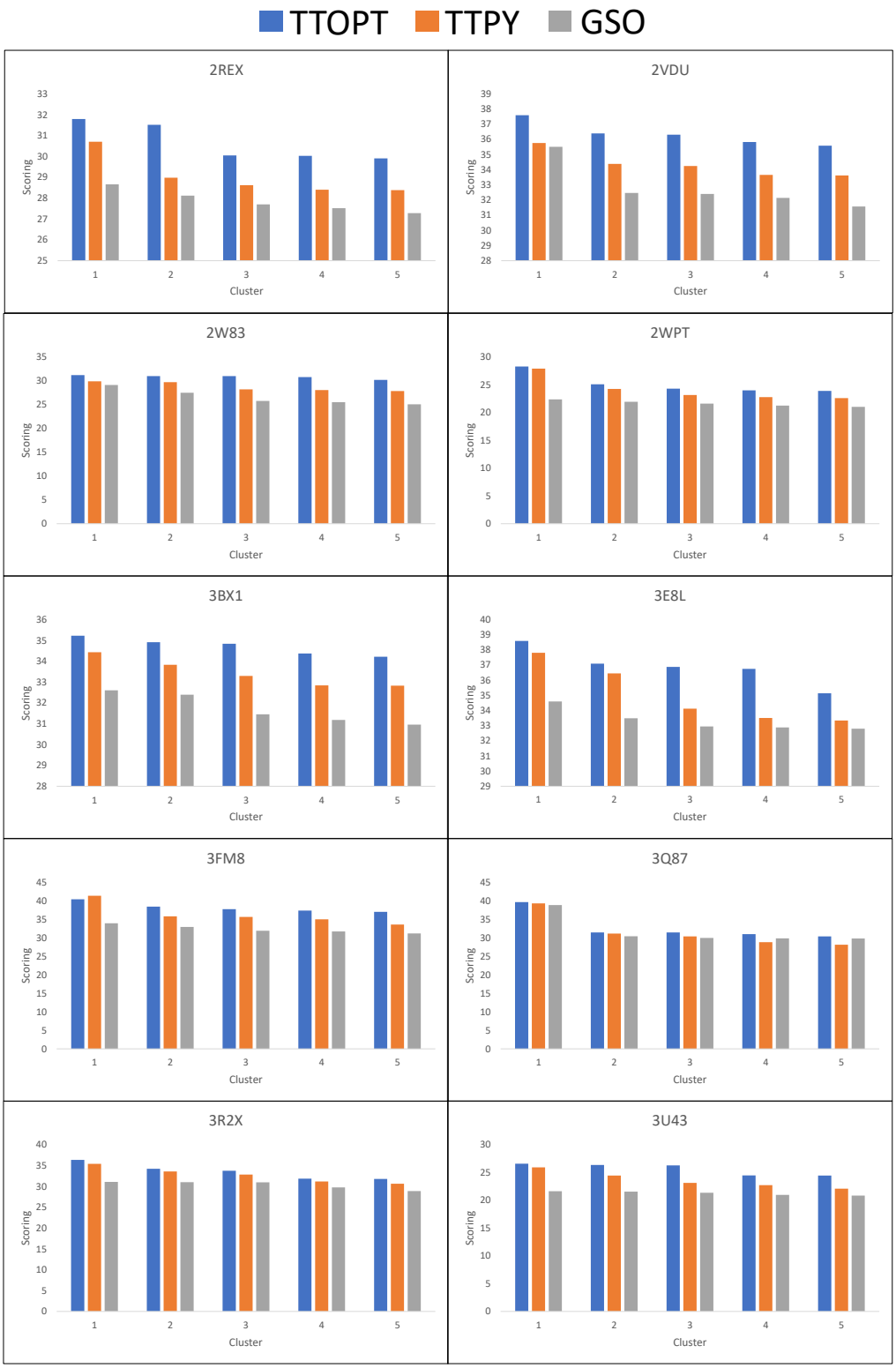

FIG. 2. Scores evaluated with the DFIRE energy function for the five top clusters of structures obtained for docking complexes with Terra Quantum's TTOPT (blue) and the TTPY (orange) and GSO (gray) approaches. Higher scores indicate more optimal solutions.

All simulations were performed in the modified version of the LightDock package that was changed in order to enable usage of the all three global minima search methods.

III. RESULTS AND DISCUSSION

The simulations were carried out using 32 CPU cores on the Google Cloud Platform. Afterwards, the simulations results were post-processed and clusterized. For tensor-train based methods, the positions of the Cα protein backbone atoms for the best ligand positions found in each of the 128 simulations were generated and clusterized using RMSD criteria of 4Å. For the GSO approach, the five best found structures of the receptor-ligand complex for each swarm were generated and further clusterized using the same 4Å RMSD criteria for the Cα atoms. The five best-scoring clusters were used for comparisons of each complex. The results are shown in Fig. 2. In the majority of cases, TTOPT outperforms GSO quite significantly in terms of the found global minima. It also outperforms the TTPY minimizer for most cases. However, in complex 3Q87, we observed almost the same results for all three methods. Inspection of the found geometries for the best-scoring complexes shows that they are nearly identical as presented for the top cluster on Fig. 3a. But in the other studied systems, the tensor-train based approaches tend to find better minima than GSO as shown for the 2REX system in Fig. 3b.

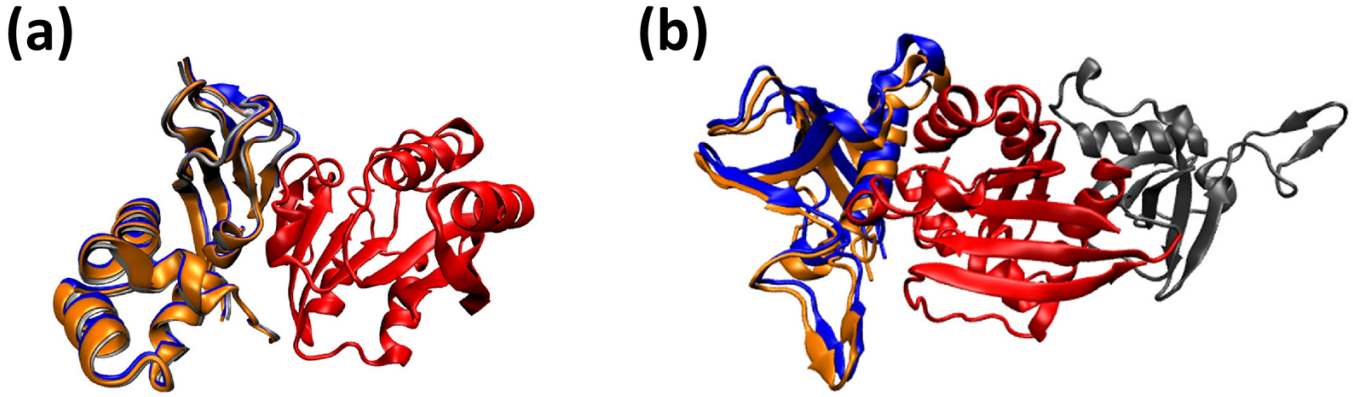

FIG. 3. Overlap of the best structures for the docking ligand protein found by employing the GSO (gray), TTOPT (blue) and TTPY (orange) algorithms. The receptor protein is shown in red. For the 3Q87 panel (a), the complexes are nearly identical. For the 2REX panel (b), the tensor-train based approaches found better minima.

IV. CONCLUSION

We have presented a novel protein–protein docking approach based on the tensor-train decomposition of multi-dimensional scoring function landscapes. Our own TTOPT algorithm was implemented in the LightDock package. It demonstrated promising capabilities for sampling the translational and rotational space of rigid body protein–protein docking simulations. We demonstrated that our method outperforms LightDock's native GSO algorithm as well as the tensor-train based black-box global optimizer implemented in the ttpy library. Due to the nature of our optimization approach, it could also be employed for other chemical and biological modeling problems, such as the folding of proteins and nucleic acids. More broadly, it could be used in machine learning pipelines and for the optimization of hyperparameters [25], for instance in the prediction of drug responses [26].

V. ACKNOWLEDGMENTS

We are grateful to Prof. V. Vinokur and Dr. Y. Alexeev from Argonne National Laboratory for valuable discussions and suggestions.

[1] David Eisenberg, Edward M Marcotte, Ioannis Xenarios, and Todd O Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823–826, 2000.

[2] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[3] Xuan-Yu Meng, Hong-Xing Zhang, Mihaly Mezei, and Meng Cui. Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7(2):146–157, 2011.

[4] Jerome De Ruyck, Guillaume Brysbaert, Ralf Blossey, and Marc F Lensink. Molecular docking as a popular tool in drug design, an in silico travel. *Advances and applications in bioinformatics and chemistry: AABC*, 9:1, 2016.

[5] Nataraj S Pagadala, Khajamohiddin Syed, and Jack Tuszynski. Software for molecular docking: a review. *Biophysical reviews*, 9(2):91–102, 2017.

[6] Tammy Man-Kuang Cheng, Tom L Blundell, and Juan Fernandez-Recio. pydock: electrostatics and desolvation for effective scoring of rigid-body protein–protein docking. *Proteins: Structure, Function, and Bioinformatics*, 68(2):503–515, 2007.

[7] Xiaofan Li, Iain H Moal, and Paul A Bates. Detection and refinement of encounter complexes for protein–protein docking: taking account of macromolecular crowding. *Proteins: Structure, Function, and Bioinformatics*, 78(15):3189–3196, 2010.

[8] KN Krishnanand and Debasish Ghose. Glowworm swarm optimization for simultaneous capture of multiple local optima of multimodal functions. *Swarm intelligence*, 3(2):87–124, 2009.

[9] Brian Jiménez-García, Jorge Roel-Touris, Miguel Romero-Durana, Miquel Vidal, Daniel Jiménez-González, and Juan Fernández-Recio. Lightdock: a new multi-scale approach to protein–protein docking. *Bioinformatics*, 34(1):49–55, 2018.

[10] Ivan Oseledets. Tensor-train decomposition. *SIAM J. Scientific Computing*, 33:2295–2317, 01 2011.

[11] Konstantin Sozykin, Andrei Chertkov, Roman Schutski, Anh-Huy Phan, Andrzej Cichocki, and Ivan Oseledets. Ttopt: A maximum volume quantized tensor train-based optimization and its application to reinforcement learning. *arXiv preprint arXiv:2205.00293*, 2022.

[12] William Huggins, Piyush Patil, Bradley Mitchell, K Birgitta Whaley, and E Miles Stoudenmire. Towards quantum machine learning with tensor networks. *Quantum Science and technology*, 4(2):024001, 2019.

[13] Manuel S Rudolph, Jing Chen, Jacob Miller, Atithi Acharya, and Alejandro Perdomo-Ortiz. Decomposition of matrix product states into shallow quantum circuits. *arXiv preprint arXiv:2209.00595*, 2022.

[14] Mohammad Kordzanganeh, Markus Buchberger, Maxim Povolotskii, Wilhelm Fischer, Andrii Kurkin, Wilfrid Somogyi, Asel Sagingalieva, Markus Pflitsch, and Alexey Melnikov. Benchmarking simulated and physical quantum processing units using quantum and hybrid algorithms. *arXiv preprint arXiv:2211.15631*, 2022.

[15] Michael Perelshtein, Asel Sagingalieva, Karan Pinto, Vishal Shete, Alexey Pakhomchik, Artem Melnikov, Florian Neukart, Georg Gesek, Alexey Melnikov, and Valerii Vinokur. Practical application-specific advantage through hybrid quantum computing. *arXiv preprint arXiv:2205.04858*, 2022.

[16] Jorge Roel-Touris, Alexandre MJJ Bonvin, and Brian Jiménez-García. Lightdock goes information-driven. *Bioinformatics*, 36(3):950–952, 2020.

[17] Ivan Oseledets. ttpy v.1.2.1, 2022.

[18] Dmitry Zheltkov and Eugene Tyrtyshnikov. Global optimization based on TT-decomposition. *Russian Journal of Numerical Analysis and Mathematical Modelling*, 35(4):247–261, 2020.

[19] Ivan Oseledets and Eugene Tyrtyshnikov. TT-cross approximation for multidimensional arrays. *Linear Algebra and its Applications*, 432(1):70–88, 2010.

[20] Sergei Goreinov, Ivan Oseledets, Dimitry Savostyanov, Eugene Tyrtyshnikov, and Nikolay Zamarashkin. How to find a good submatrix. In *Matrix Methods: Theory, Algorithms And Applications: Dedicated to the Memory of Gene Golub*, pages 247–256. World Scientific, 2010.

[21] Micheline B Soley, Paul Bergold, and Victor S Batista. Iterative power algorithm for global optimization with quantics tensor trains. *Journal of Chemical Theory and Computation*, 17(6):3280–3291, 2021.

[22] Marc F Lensink and Shoshana J Wodak. Score_set: a capri benchmark for scoring protein complexes. *Proteins: Structure, Function, and Bioinformatics*, 82(11):3163–3169, 2014.

[23] CHI Zhang, Song Liu, and Yaoqi Zhou. Accurate and efficient loop selections by the dfire-based all-atom statistical potential. *Protein science*, 13(2):391–399, 2004.

[24] Michael JD Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, 7(2):155–162, 1964.

[25] Asel Sagingalieva, Andrii Kurkin, Artem Melnikov, Daniil Kuhmistrov, Michael Perelshtein, Alexey Melnikov, Andrea Skolik, and David Von Dollen. Hyperparameter optimization of hybrid quantum neural networks for car classification. *arXiv preprint arXiv:2205.04878*, 2022.

[26] Asel Sagingalieva, Mohammad Kordzanganeh, Nurbolat Kenbayev, Daria Kosichkina, Tatiana Tomashuk, and Alexey Melnikov. Hybrid quantum neural network for drug response prediction. *arXiv preprint arXiv:2211.05777*, 2022.