Scalable prediction of symmetric protein complex structures

Victor Yu[1, *], Perry Demsko[1], Roger Castells-Graells[2], Hank Parker[1], Andrew Huang[1], Chloe Chen[1], Martin Huang[1], Varsitaa Srinivasan[1], Krishna Ajjarapu[1], Nica Tofighbakhsh[1], Richard Yu[1], Michael Lake[3, 4], David L. Glanzman[5, 6, 7], Sarah Warren[8], and Joseph Alzagatiti[1, *]

[1]EndoFold, Los Angeles, CA, 90095, USA
[2]Structural Biology Programme, Spanish National Cancer Research Centre (CNIO), Madrid, 28029 Spain
[3]California NanoSystems Institute, University of California, Los Angeles, Los Angeles, CA 90095 USA
[4]BioPACIFIC Materials Innovation Platform Living Biofoundry, University of California, Los Angeles, Los Angeles, CA 90095 USA
[5]Department of Integrative Biology and Physiology, University of California, Los Angeles, Los Angeles, CA, 90095, USA
[6]Department of Neurobiology, David Geffen School of Medicine at UCLA, Los Angeles, CA, 90095, USA
[7]Integrative Center for Learning and Memory, Brain Research Institute, David Geffen School of Medicine at UCLA, Los Angeles, CA, 90095, USA
[8]Gilead Sciences, Seattle, WA 98105 USA
[*]Contributed Equally; Corresponding Authors: Victor Yu, victor.yu@endofold.bio and Joseph Alzagatiti, joseph.alzagatiti@endofold.bio

November 15, 2025

Abstract

All life relies on proteins to function, yet accurately modeling protein structures exceeding 10,000 amino acids remains extremely difficult. Existing solutions are limited to specific scenarios, require considerable computational resources, or are otherwise unscalable. Consequently, many large, disease-relevant protein complexes in the human proteome, as well as nearly all viruses and numerous other classes, are impractical to model with high fidelity for drug development. To modulate these protein complexes and viruses, structural information is eminently valuable, and often essential. In the last two years, machine learning based-tools that can generate binders to a given target structure with high hit rates have emerged. Combined with high-throughput screening, these technologies can far outpace traditional drug discovery. However, they cannot function well without accurate models of their target structures. Thus, to unlock the full power of AI-driven drug discovery, a scalable method must be developed to predict large protein complex structures. To overcome this bottleneck, we introduce Cosmohedra, a physics-based method to rapidly and accurately predict the structure of arbitrarily large, symmetric protein complexes. Validated across 4 major symmetry classes (icosahedral, tetrahedral, octahedral, and cyclic), the method consistently achieves near-experimental levels of accuracy, i.e., RMSD < 5Å. In test cases, the method runs in < 5 minutes on consumer hardware, 10^3-10^5 times faster than the closest comparable software. The largest structure currently built, at ≈40,000 amino acids, is > 4 times the limits of existing machine learning and molecular dynamics-based methods. By dramatically increasing the speed and scale at which protein complex structures can be modeled, Cosmohedra represents a new step towards universal protein structure prediction and a valuable tool for protein engineering and drug development.

1 Introduction and motivation

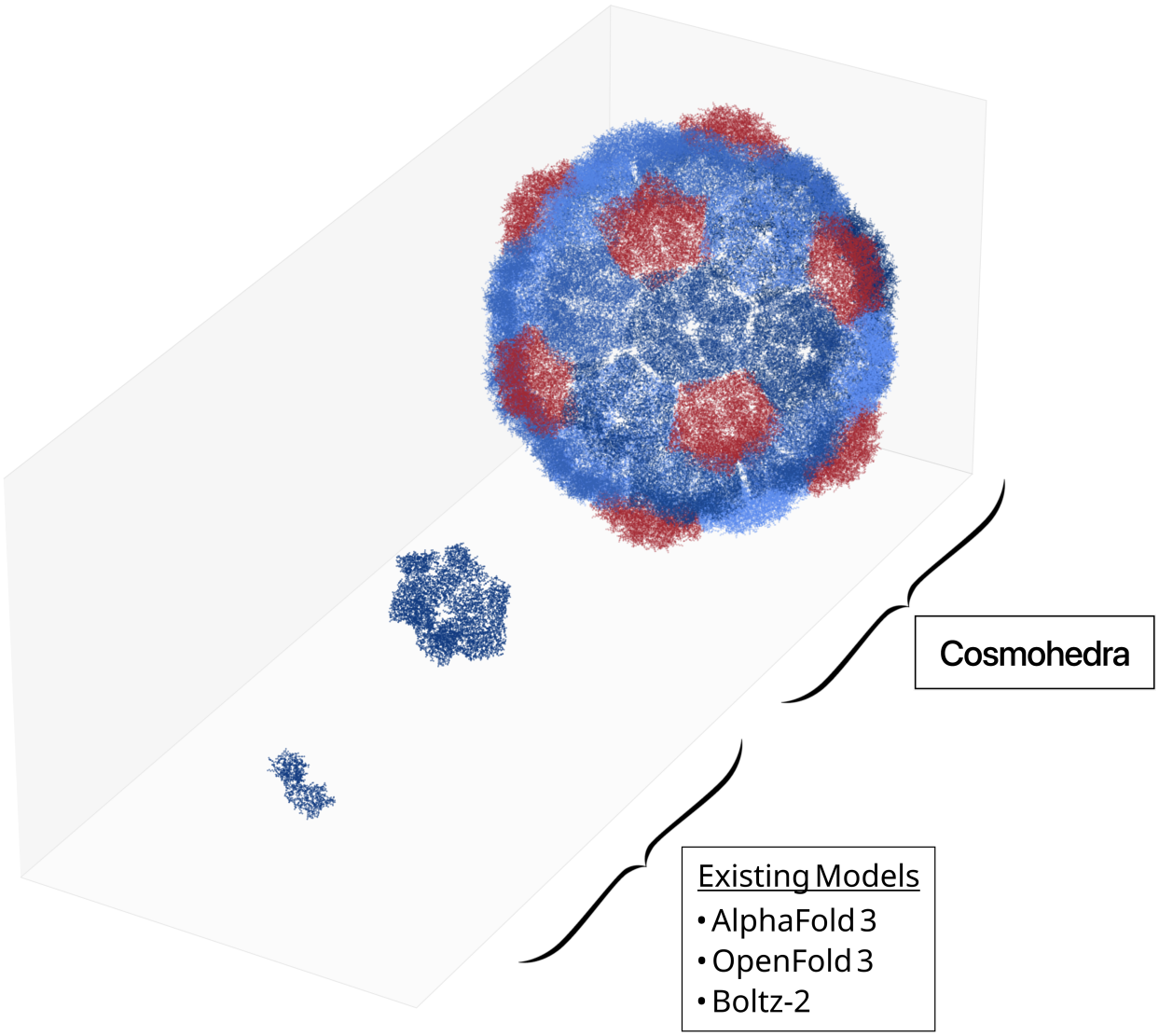

Figure 1: *From left to right*, the monomer (modeled with AlphaFold2), multimer (modeled with AlphaFold-Multimer), and complex (modeled with Cosmohedra) level structures of the dARC2 capsid of *Drosophila melanogaster* [1].

Developing a fast and accurate algorithm that can predict the structure of any protein assembly solely from its amino acid sequence is a central goal of structural biology.

Protein structures exist at three levels of scale: monomer, multimer, and complex (**Figure 1**). From the perspective of a protein complex, a single monomer corresponds to a brick; a single constituent multimer corresponds to a wall; and the entire complex a house. Within *in vivo* systems, components of large protein complexes rarely assemble simultaneously, but rather in a sequential manner. Complexes often assemble monomers to multimers before fitting these multimers into the final assembly, prioritizing optimal interfaces between protein monomers before subsequent chains can be added [2, 3].

After decades of incremental progress, accurate sequence-to-structure prediction was accomplished by AlphaFold2 at the monomer-level in 2020, revolutionizing structure-based drug design and discovery [4].

Soon, AlphaFold2 reached the multimer-level in 2021, opening the door to modeling and designing protein-protein interactions and small complexes [5]. The implications for protein engineering, pharmaceuticals, and medical research in general are paradigm-altering [6].

Unfortunately, progress toward predicting structures at the protein complex level has stalled due to two fundamental issues: GPU memory constraints and the lack of experimental data on large protein complexes.

The transformer-style attention mechanisms that power AlphaFold and many other machine learning models inherently run in $O(n^2)$ time [4, 5, 7, 8]. Although there are approximate or restricted attention mechanisms that operate closer to linear time, performance may be sacrificed, i.e., the native sparse attention basis of DeepSeek V3.2 [9]. The quadratic scaling of these models makes it extremely difficult to build larger complexes since the memory bandwidth limit on GPUs is quickly exceeded [10]. AlphaFold3 switched to a diffusion architecture partly to compensate for the memory efficiency concern [8]. However, despite this architecture being more memory efficient, it is less compatible with parallelization operations. Altogether, this means AlphaFold3 has a size limit at ≈ 5000 amino acids, beyond which accuracy drops precipitously.

The more intractable issue is data availability. Essentially all publicly available protein structural databases in the world have already been ingested by the current AlphaFold models [4, 5, 8]. To enable progress toward large protein complexes and other important areas like antibody-receptor docking, new structural data is needed. This comes primarily from two sources: Cryo-electron microscopy (cryoEM) and X-ray diffraction (XRD), both of which are time-consuming, technically difficult, and resource-intensive procedures. These techniques can take weeks to months, even in the hands of an experienced operator, but sample purification and isolation success are also heavily dependent on the protein target of interest [11]. There are likely fewer than 1000 cryoEM systems in the United States due to the significant infrastructure investment required [12]. Even if operated simultaneously at full capacity for a year, they would likely image far less than the tens of thousands of complexes required for training a new AlphaFold-like model. Experimentally generating sufficient quantities of data on large protein complexes to train new AlphaFold-like models is currently infeasible.

In the generative AI era, protein design tools such as Latent-X, RFDiffusion3, and BoltzGen are widely accessible to the community [13, 14, 15]. When provided with clear target structures, these tools, combined with the emerging rapid screening capabilities of cloud labs, will outpace traditional technologies like phage display and animal antigen immunization for antibody identification [16]. They are now bottlenecked by target discovery, i.e., selecting what proteins and where on the protein structure to target. Much of the human proteome, as well as those of viruses, animals, and plants, remains undruggable because we cannot model its structures. Therefore, there is a crucial need for modeling large protein complexes (i.e. > 60 subunits and $> 10,000$ amino acids) at computational scale in order to identify key drug targets.

With our model, which we name Cosmohedra, we set out to circumvent the limitations imposed by transformer architectures and experimental data availability by using physics-based heuristics. The use of such heuristics and specific internal data structures allow Cosmohedra to run in $O(n \log n)$ time without sacrificing performance. It is also parallelizable and is not limited by hardware memory capacity. As a result, it is able to model protein complex structures far above 10,000 amino acids, with the largest demonstration at $\approx$40,000.

2 The Cosmohedra model

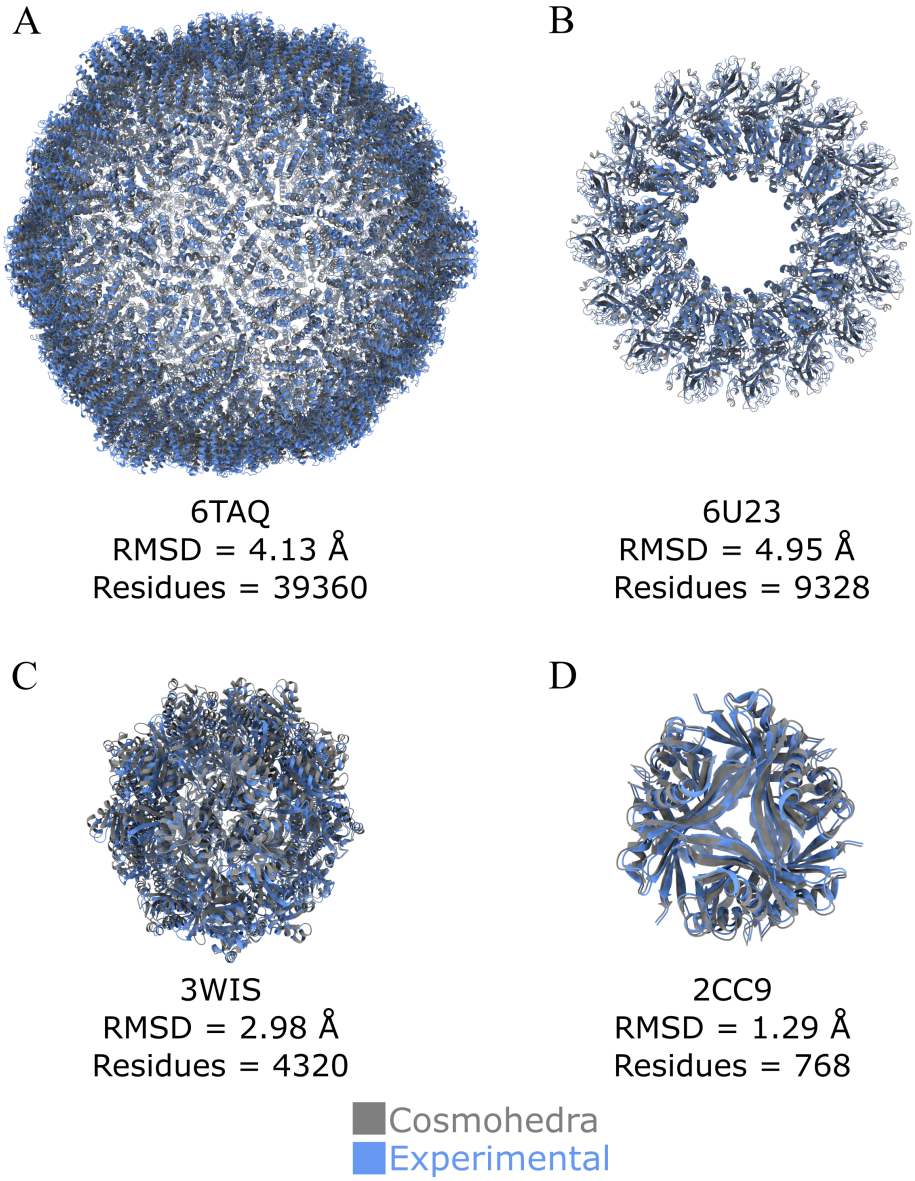

Figure 2: Cosmohedra predicted protein complex structures overlapping with known cryoEM or X-ray crystallography data in the Protein Data Bank (PDB) [17]. **A)** The dARC2 capsid of *Drosophila melanogaster*. This structure is $> 40,000$ amino acids in size, but the number of modeled residues are truncated in the cryoEM resolved structure. Iscosahedral symmetry complex. PDB code: 6TAQ [1] **B)** Human macrophage-expressed gene 1. Cyclic symmetry. PDB code: 6U23 [18] **C)** Protein cubic cage for redox transfer in *Paraburkholderia xenovorans*. Octahedral symmetry complex. PDB code: 3WIS [19] **D)** Dodecin in *Halobacterium salinarum R1*. Tetrahedral symmetry complex. PDB code: 2CC9 [20]

To build a complex, Cosmohedra is supplied with its symmetry type and predicted subunit structure, generated by an AlphaFold-like model. For example, for the complex in Figure 2C, Cosmohedra is given its octahedral symmetry and its predicted multimer structure (in this case, a trimer). Cosmohedra then assembles instances of the trimer into a cohesive complex, structuring interfaces to optimize internal parameters. The completed structure is then outputted.

Existing "docking" methods use a similar approach [21, 22, 23]. Such methods, whether structure-based, machine learning-based, or physics-based, serve a crucial purpose but have certain limitations. Some structure-based methods, such as RPXdock, are relatively fast and are sequence-independent [21]. This method performs best with experimentally verified or previously designed inputs. As such, publicly available literature seems to primarily focus on synthetic designs instead of naturally-occurring protein complexes, which may require more experimental data in order to model [24, 25]. A machine learning-based method, CombFold, uses AlphaFold-Multimer for combinatorial assembly of homomeric and heteromeric protein complexes [22]. This method is limited by AlphaFold-Multimer's ability to produce pariwise subunit interactions, which, while state-of-the-art, is limited by the availability of training data, especially of protein complexes. It also has relatively long runtimes, up to ≈ 250 minutes in test cases, and a size limitation of 30 protein chains or 18,000 amino acids [22]. Some prominent methods are physics-based and rely on molecular dynamics calculations. These are accurate and versatile, creating structures with accuracies on par with experimental data resolutions. The most notable of these is EvoDock, which is the nearest comparable technology to Cosmohedra and is the current state-of-the-art [23]. EvoDock achieves extremely high accuracies (often better than Cosmohedra), but requires tens to hundreds of hours of GPU time for a detailed simulation of a single, small complex, and has a $\approx 10^4$ amino acid size limit.

Cosmohedra seeks to address all of the limitations discussed above with specific protein prediction models and docking methods. The use of physics heuristics instead of transformer architecture or molecular dynamics calculations makes Cosmohedra extremely efficient. For these reasons, it is scalable to arbitrarily large complexes. The largest structure constructed is $> 40,000$ amino acids and over 4 MegaDaltons (**Figure 1; 2A**). This exceeds the currently observed limits of molecular dynamics docking methods such as EvoDock by 4-fold and protein prediction models like AlphaFold3 by 8-fold. We eschew existing symmetry definitions used by many docking tools in favor of deriving our own symmetry library using geometric methods. This enables the modeling of naturally occurring higher-order ($\mathrm{T} > 1$) icosahedral and large cyclic complexes. To our knowledge, this work represents the first fast and accurate predictions of our test cases — the human macrophage-expressed gene 1 (PDB code: 6U23) and the dARC2 capsid of *Drosophila melanogaster* (PDB code: 6TAQ) — without experimental data templates and extensive molecular dynamics calculations [1, 18].

To demonstrate the model's capabilities, we built 36 complexes from sequence, and compared them to their experimentally measured structures in the Protein Data Bank (PDB) (**Figures 2; 3**); see appendix for a table of the complexes and their accuracies. Of note, nearly all predictions had accuracies < 5Å root mean square deviation (RMSD) compared to the experimental data, which makes the structures sufficiently accurate for drug design [26]. The two outliers had errors associated with the symmetry libraries which have since been identified and are currently being resolved. Using inaccurate AlphaFold-like predictions as inputs can damage the accuracy; therefore, all subunit structures used in this work had RMSD < 4Å.

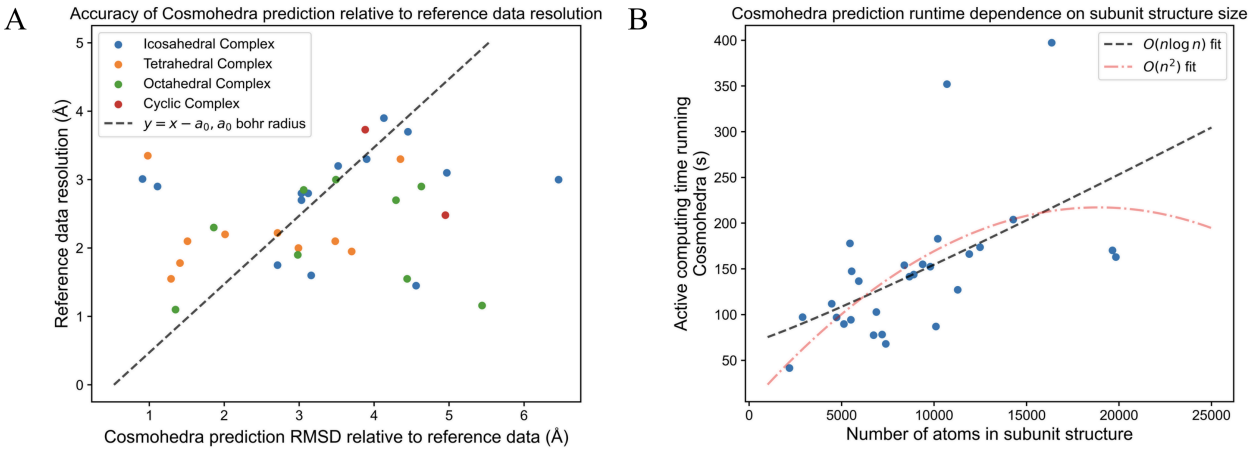

Figure 3: **A)** Accuracy of Cosmohedra predictions with respect to cryoEM and X-ray crystallography reference data, plotted against the internal resolution of that data. **B)** Cosmohedra prediction runtime plotted against the number of atoms in the subunit structure of the predicted complex, demonstrating an $O(n \log n)$ trend.

About half of the predictions have accuracies within one bohr radius of the resolution of the benchmark experimental data or better (**Figure 3A**). This suggests that the model is approaching the noise floor of the experimental methods used to generate the reference data. Beyond this level, a lower RMSD holds diminished meaning. The accuracy of the reference data is subject to fundamental physical limits and a more accurate experimental benchmark has yet to be invented [27, 28].

Finally, we note that Cosmohedra's required runtime for these predictions follows the expected $O(n \log n)$ trend (**Figure 3B**). An $O(n^2)$ trend, as might be expected in some of the previously mentioned methods, cannot be fitted to the data without incorrectly suggesting that computing time decreases with increasing subunit size, and indicates that Cosmohedra does not run in $O(n^2)$ time. The timed computations are for complexes featured in the EvoDock method's publication. The same complexes which ran for tens to hundreds of hours on EvoDock were constructed by Cosmohedra in minutes [23]. The Cosmohedra runtimes recorded in Figure 3 were performed locally on a MacBook Air with a Apple M2 chip. Cosmohedra has also been successfully parallelized to run on cloud-based virtual machines. A 96-core CPU can perform these same calculations in roughly 5-30 seconds each. This makes Cosmohedra 10^3 to 10^5 times faster than current state-of-the-art methods.

We note that the cyclic complexes and the T = 4 icosahedral capsid's computing times were not included in Figure 3B. These complexes require a higher dimension parameter space and run in about 20-30 minutes on a 96-core CPU, but still run on $O(n \log n)$ time.

3 Predicting structures outside of the PDB and building a general antiviral engine

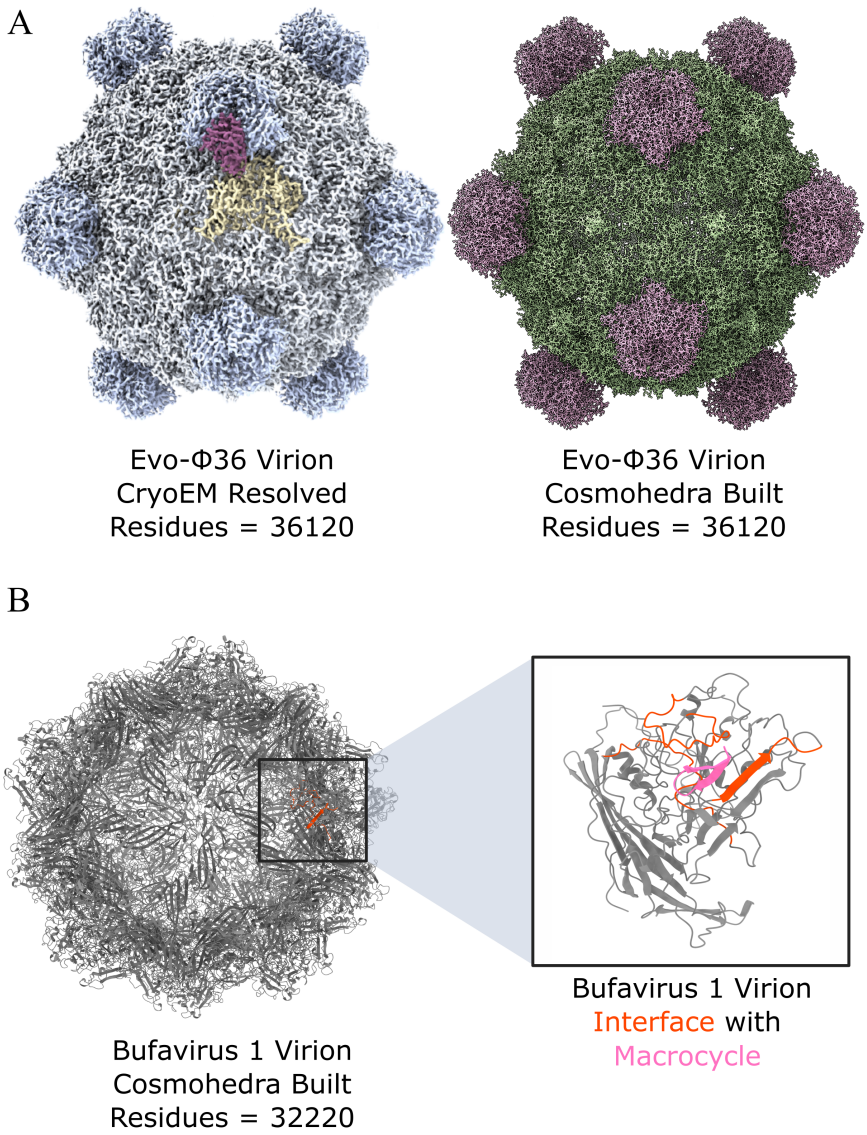

Figure 4: **A)** *Left*, Image of Evo-Φ36's cryoEM structure, reproduced from Arc Institute publication [29]. *Right*, Cosmohedra's prediction of Evo-Φ36 from its publicly-released genome. **B)** *Left*, Cosmohedra prediction of Bufavirus 1 capsid structure, $6-7$Å RMSD relative to known cryoEM datasets. *Right*, Macrocycle inhibitor bound to Bufavirus 1 capsid, designed with Latent-X.

Having predicted known structures with greatly reduced computation times and with high fidelity using Cosmohedra, we next wanted to demonstrate the ability to predict structures not seen in the PDB, including those designed by large machine learning models. We chose Evo-Φ36, a bacteriophage designed with the Evo2 large language model, as our proof-of-concept task [29]. At the time of publication, it represented the first functional, AI-generated genome and did not have a publicly available structure in the PDB to reference. The publicly released genome and homology observations were used to annotate the Capsid F and Spike G protein of the AI designed phage. The joint pentameric structure were then predicted in AlphaFold3 (**Figure 4A**). Cosmohedra then assembled the structure into a full, 3D complex with striking accuracy compared to

the 2D representation of the CryoEM structure (**Figure 4A**), and $\approx$ 4Å RMSD relative to the wildtype phage Evo-Φ36 was engineered from. Notably, this complex is heteromeric, demonstrating that Cosmohedra is not limited to homomeric complex assemblies. Prediction of the Evo-Φ36 virus points to the application of Cosmohedra towards rapid antiviral design.

High resolution structural data on a virus can enable, or at least greatly facilitate, the rational design of specific antivirals targeting that virus [30, 31, 32]. New tools, such as Latent-X, RFDiffusion3, and BoltzGen, are capable of robust, semi-automated inhibitor design against protein complexes given their structure. Once a virus's structure has been determined, antivirals can be designed against it. Unfortunately, virus structures are historically difficult to determine experimentally. In addition to all the usual challenges of cryoEM and XRD, viral pathogens work often requires a BSL-3 or even BSL-4 facility [33]. This compounded challenge often means that viral structural determination takes at least one month or, more commonly, multiple months. By the time the structure of an emerging pathogen has been experimentally determined, an outbreak may have already spread.

The most efficient method would be to predict a virus's structure from its sequence purely computationally. This is impractical with available tools [34], but we believe Cosmohedra bridges this gap. To demonstrate this capability, we used Cosmohedra to predict the Bufavirus 1 capsid structure and identify sites of interaction for targeting with macrocycles to inhibit capsid formation (**Figure 4B**). Using Latent-X, we generated $\approx$ 300 candidates. *In vitro* validation is underway.

Combining the rapid protein complex prediction capabilities of Cosmohedra with structure-based generative AI drug design models and high-throughput screening, we form the basis of a general antiviral engine. Using this workflow, a library of antiviral candidates can be designed beginning with viral sequence in under 3 hours. If sufficiently validated, this would be paradigm-altering for antiviral development. The Bufavirus 1 was chosen because it is difficult to model with Cosmohedra ($6-7$Å RMSD relative to experimental data) while three high quality cryoEM structures have been reported at different pH levels [35]. This provides a reference point to help understand how AI-designed binders interact with the Cosmohedra-predicted structure, and how the workflow may be improved. Crucially, Bufavirus 1, to our knowledge, has no specific antiviral. A successful deployment of an antiviral in a real-world setting against Bufavirus 1 would thus be a definitive proof-of-concept.

As demonstrated painfully by the COVID-19 pandemic, essentially no country can guarantee its ability to respond quickly enough to contain an outbreak without the risk of catastrophic socioeconomic consequences. A warming and more interconnected world is at increasing risk for more global pandemics [36, 37, 38]. Beyond these naturally evolving risks, the emergence of AI-designed viruses, such as Evo-Φ36, poses an entirely new threat. Tools for designing viruses have far outpaced biosecurity infrastructure and legislation, even before the LLM era [39]. Open-access and open-source practices for leading Bio-AI tools have removed significant barriers for developing dual-use biotechnologies [40]. For both global pandemic preparedness and biodefense, there is urgent need for a universal antiviral engine.

The Cosmohedra-based antiviral engine is currently built for icosahedral viruses, which represent about half of known viruses [41, 42, 43]. To bridge the gap toward a universal antiviral engine, we are in the process of enabling other symmetry classes, namely helical viruses. A more difficult proposition is providing foundational models with the data they need to better model both natural and artificial virus multimers for Cosmohedra to assemble and for our antiviral engine to target. A similar argument applies to the path towards universal protein structure prediction and protein complex modulation.

4 Toward universal protein structure prediction and protein complex modulation

Beyond viruses, we believe Cosmohedra's ability to generate protein complexes *de novo* allows it to create an unrivaled synthetic dataset for training next-generation foundational models. Chai-2 demonstrated material improvement in the field of antibody-antigen engineering with the addition of synthetic data [44]. Further development of Cosmohedra could provide data across protein complex classes for the next generation of models to learn. Essentially, it could create a synthetic counterpart to the PDB. This moves us towards a more complete foundational model of protein biology.

A limited, proof-of-concept version can be used to strengthen our technology for pandemic preparedness and biodefense. While AlphaFold3's multimer predictions are state-of-the-art, they often fail for structures with insufficient homologous proteins in the PDB [45, 46]. By fine-tuning an open-source, AlphaFold-like model on a synthetic dataset of capsid complexes, we hope to improve upon capsid protein predictions.

We can then extend the model to general human health. The human proteome contains approximately 60 to 70 percent symmetric protein complexes [47]. However, the current human protein complex map, hu.MAP 2.0, has a set of ∼5000 complexes with more than 2 chains, and only ∼400 of these complexes have any reported structures [23, 48]. Complete structures exist for only a small subset of human complexes like the proteasome, ribsome, EIF2B complex, and spliceosome [48]. Furthermore, interface regions of human protein complexes show a significant enrichment of disease-causing mutations [47]. Therefore, by moving closer to modeling the entire human proteome, especially protein complex interfaces, we can unlock a substantial set of novel, druggable targets. This would be a significant step towards the eventual, field-wide goal of universal protein complex modulation: the ability to make a drug targeting any protein complex in any species by modeling its structure and designing against it. The basis of that technology would have to involve scalable prediction of large protein complex structures, and likely the synthetic data that can subsequently be generated.

In this context, it is also important to note Cosmohedra's limitations and associated future directions of work. Currently, Cosmohedra requires both the monomer sequence and symmetry of a complex, and models only symmetric protein complexes. While we see a clear path to important classes such as antibody-antigen complexes, there are many asymmetric complexes Cosmohedra cannot model. Cosmohedra has not yet been used to performed an end-to-end blind prediction: modeling a complex solely from sequence and symmetry when no experimental image of any kind has been generated. Such efforts are underway — a cryoEM image of a capsid structure is being produced, with the authors predicting its structure using Cosmohedra in isolation from any experimental data. We also note that the entire Cosmohedra workflow begins with high quality protein sequence. While the complex's symmetry can often be deduced from the sequence, this is not always the case. Relying on sequence also makes the antiviral applications reactive. It cannot be deployed in advance of emerging viral threats as a preventative measure.

Meanwhile, new approaches to structural prediction that underpin Cosmohedra offer opportunities for further exploration and development. With those advancements, we hope aforementioned limitations can be overcome, and new application areas discovered.

5 Conclusion

Cosmohedra enables *de novo* prediction of large protein complex structures beyond the scale and limits of current machine learning prediction and molecular-dynamics docking methods. Operable on consumer hardware, e.g. a MacBook Air with a Apple M2 chip, it can run near-experimental resolution predictions in minutes.

Validation across symmetry classes against known structures, as well as the prediction of unknown structures, demonstrates the capability of this method in protein engineering and drug development. The ability to generate large complexes quickly, without specialized hardware, can enable applications like rapid antiviral design and generating data for training future foundational protein models.

We hope Cosmohedra will serve as a valuable tool for structural biology and AI-driven drug discovery, and help move the field closer toward universal protein structure prediction and protein complex modulation.

Acknowledgement

The authors would like to thank Rebecca Allan, Colin Maraganore, and Nelly Weiser for helpful discussions on the manuscript. We thank Seshu Ajjarapu for his mentorship and guidance. We are grateful to the facilities and operations staff at the California NanoSystems Institute for maintaining a safe and productive work space. We also acknowledge Jackson Chen for help in generating the figures in this work.

References

[1] Sigurbjörn Erlendsson, Dustin R. Morado, Heather B. Cullen, Cédric Feschotte, Jason D. Shepherd, and John A. G. Briggs. Structures of virus-like capsids formed by the Drosophila neuronal Arc proteins. *Nature Neuroscience*, 23:172–175, 2020.

[2] P. Bryant et al. Predicting the structure of large protein complexes using alphafold and monte carlo tree search. *Nature Communications*, 13:6028, 2022.

[3] Joseph A. Marsh, Helena Hernández, Zoe Hall, Sebastian E. Ahnert, Tina Perica, Carol V. Robinson, and Sarah A. Teichmann. Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell*, 153(2):461–470, April 2013.

[4] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Jul 2021.

[5] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *bioRxiv*, Oct 2021.

[6] The Royal Swedish Academy of Sciences. The nobel prize in chemistry 2024. Press release, October 2024.

[7] Mehmet Akdel, Douglas E. Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O. Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L. Good, Roman A. Laskowski, Gabriele Pozzati, et al. A structural biology community assessment of alphafold2 applications. *Nature Structural and Molecular Biology*, 29(11):1056–1067, Nov 2022.

[8] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, May 2024.

[9] DeepSeek AI. Deepseek-v3.2-exp. https://huggingface.co/deepseek-ai/DeepSeek-V3.2-Exp, 2025. Accessed: 2025.

[10] M. Mirdita, K. Schütze, Y. Moriwaki, et al. Colabfold: making protein folding accessible to all. *Nature Methods*, 19:679–682, 2022.

[11] Bruce Alberts. The cell as a collection of protein machines: Preparing the next generation of molecular biologists. *Cell*, 92(3):291–294, Feb 1998.

[12] Creative Biostructure. Comparison of x-ray crystallography, nmr and em. *Technical Report*, 2025.

[13] Latent Labs. Latent labs platform. https://platform.latentlabs.com/, 2025. Accessed: 2025.

[14] Jasper Butcher, Rohith Krishna, Raktim Mitra, Rafael I. Brent, Yanjing Li, Nathaniel Corley, Paul Kim, Jonathan Funk, Simon Mathis, Saman Salike, et al. De novo design of all-atom biomolecular interactions with rfdiffusion3. *bioRxiv*, Sep 2025.

[15] Hannes Stark and BoltzGen Team. Boltzgen: All-atom generative diffusion for universal binder design. https://boltz.bio/boltzgen, 2025. MIT License. Accessed: 2025.

[16] Carrie Arnold. Cloud labs: Where robots do the research. *Nature*, 606(7914):612–613, Jun 2022.

[17] Stephen K. Burley, Helen M. Berman, Gerard J. Kleywegt, John L. Markley, Haruki Nakamura, and Sameer Velankar. Protein data bank (pdb): The single global macromolecular structure archive. *Methods in Molecular Biology*, 1607:627–641, 2017.

[18] Stephanie S. Pang, Charles Bayly-Jones, Mazdak Radjainia, Bradley A. Spicer, Ruby H. P. Law, Amélie W. Hodel, Emma S. Parsons, Sarah M. Ekkel, Phillip J. Conroy, Georg Ramm, James C. Whisstock, Michelle A. Dunstone, and Ilia Voskoboinik. The cryo-EM structure of the acid activatable pore-forming immune effector macrophage-expressed gene 1. *Nature Communications*, 10:4288, 2019.

[19] Dan E. McNamara, Duilio Cascio, Julien Jorda, Cheene Bustos, Tzu-Chi Wang, Madeline E. Rasche, Todd O. Yeates, and Thomas A. Bobik. Structure of dihydromethanopterin reductase, a cubic protein cage for redox transfer. *Journal of Biological Chemistry*, 289(13):8852–8864, Mar 2014.

[20] Martin Grininger, Kornelius Zeth, and Dieter Oesterhelt. Dodecins: A family of lumichrome binding proteins. *Journal of Molecular Biology*, 357(3):842–857, Mar 2006.

[21] W. Sheffler, E. C. Yang, Q. Dowling, Y. Hsia, C. N. Fries, J. Stanislaw, et al. Fast and versatile sequence-independent protein docking for nanomaterials design using rpxdock. *PLoS Computational Biology*, 19(5):e1010680, May 2023.

[22] Ben Shor and Dina Schneidman-Duhovny. Combfold: predicting structures of large protein assemblies using a combinatorial assembly algorithm and alphafold2. *Nature Methods*, 21:477–487, 2024.

[23] Mads Jeppesen and Ingemar André. Accurate prediction of protein assembly structure by combining alphafold and symmetrical docking. *Nature Communications*, 14(1), Dec 2023.

[24] Quinton M. Dowling, Young-Jun Park, Chelsea N. Fries, Neil C. Gerstenmaier, Sebastian Ols, Erin C. Yang, Adam J. Wargacki, Annie Dosey, Yang Hsia, Rashmi Ravichandran, Carl D. Walkey, Anika L. Burrell, David Veesler, David Baker, and Neil P. King. Hierarchical design of pseudosymmetric protein nanocages. *Nature*, 638:553–561, Feb 2025.

[25] Sangmin Lee, Ryan D. Kibler, Green Ahn, Yang Hsia, Andrew J. Borst, Annika Philomin, Madison A. Kennedy, Buwei Huang, Barry Stoddard, and David Baker. Four-component protein nanocages designed by programmed symmetry breaking. *Nature*, 638:546–552, Feb 2025.

[26] Hannes Stark, Felix Faltings, MinGyu Choi, Yuxin Xie, Eunsu Hur, Timothy O'Donnell, Anton Bushuiev, Talip Uçar, Saro Passaro, Weian Mao, Mateo Reveiz, Roman Bushuiev, Tomáš Pluskal, Josef Sivic, Karsten Kreis, Arash Vahdat, Shamayeeta Ray, Jonathan T. Goldstein, Andrew Savinov, Jacob A. Hambalek, Anshika Gupta, Diego A. Taquiri-Diaz, Yaotian Zhang, A. Katherine Hatstat, Angelika Arada, Nam Hyeong Kim, Ethel Tackie-Yarboi, Dylan Boselli, Lee Schnaider, Chang C. Liu, Gene-Wei Li, Denes Hnisz, David M. Sabatini, William F. DeGrado, Jeremy Wohlwend, Gabriele Corso, Regina Barzilay, and Tommi Jaakkola. BoltzGen: Toward universal binder design. 2025. Preprint.

[27] Anling Cheng, Jordan Hummel, Hui Qiu, and Martin Caffrey. The resolution in x-ray crystallography and single-particle cryogenic electron microscopy. *Crystals*, 10(7):580, 2020.

[28] Richard Henderson. The potential and limitations of neutrons, electrons and x-rays for atomic resolution microscopy of unstained biological molecules. *Quarterly Reviews of Biophysics*, 28(2):171–193, 1995.

[29] S. King et al. Generative design of novel bacteriophages with genome language models. *bioRxiv*, 2025. Preprint.

[30] Erik De Clercq and Guangdi Li. Antiviral agents: structural basis of action and rational design. *Clinical Microbiology Reviews*, 26(3):595–625, 2013.

[31] Erik De Clercq. Strategies in the design of antiviral drugs. *Nature Reviews Drug Discovery*, 1(1):13–25, 2002.

[32] Robert L. Grime, Martin Zoltner, Carla M. Swanson, and Erik De Clercq. Virus structure and structure-based antivirals. *Current Opinion in Virology*, 50:22–30, 2021.

[33] Wei Zhang and Stephen C. Harrison. Structural biology of viral pathogens in high-containment facilities. *Current Opinion in Structural Biology*, 52:62–69, 2018.

[34] Juan R. Perilla, Jodi A. Hadden, Boon Chong Goh, Christopher G. Mayne, and Klaus Schulten. All-atom molecular dynamics of virus capsids as drug targets. *The Journal of Physical Chemistry Letters*, 7(10):1836–1844, 2016.

[35] M. Gulkis, M. Luo, P. Chipman, M. Mietzsch, M. Söderlund-Venermo, A. Bennett, and R. McKenna. Structural characterization of human bufavirus 1: Receptor binding and endosomal ph-induced changes. *Viruses*, 16(8):1258, 2024.

[36] Katherine F. Smith, Michael Goldberg, Samantha Rosenthal, Lynn Carlson, Jane Chen, Cici Chen, and Sohini Ramachandran. Global rise in human infectious disease outbreaks. *Journal of the Royal Society Interface*, 11(101):20140950, December 2014.

[37] Kate E. Jones, Nikkita G. Patel, Marc A. Levy, Adam Storeygard, Deborah Balk, John L. Gittleman, and Peter Daszak. Global trends in emerging infectious diseases. *Nature*, 451(7181):990–993, February 2008.

[38] Christine K. Johnson, Peta L. Hitchens, Pranav S. Pandit, Julie Rushmore, Tierra S. Evans, Cristin C. W. Young, and Megan M. Doyle. Global shifts in mammalian population trends reveal key predictors of virus spillover risk. *Proceedings of the Royal Society B: Biological Sciences*, 287(1924):20192736, April 2020.

[39] National Academies of Sciences, Engineering, and Medicine. *Safeguarding the Bioeconomy*. The National Academies Press, Washington, DC, 2020.

[40] Nuclear Threat Initiative. The convergence of artificial intelligence and the life sciences: Safeguarding technology, rethinking governance, and preventing catastrophe, October 2023.

[41] Reidun Twarock and Antoni Luque. Structural puzzles in virology solved with an overarching icosahedral design principle. *Nature Communications*, 10(1):4414, Sep 2019.

[42] T. S. Baker, N. H. Olson, and S. D. Fuller. Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. *Microbiology and Molecular Biology Reviews*, 63(4):862–922, 1999.

[43] J. E. Johnson and J. A. Speir. Quasi-equivalent viruses: a paradigm for protein assemblies. *Journal of Molecular Biology*, 269:665–675, 1997.

[44] Chai Discovery Team. Zero-shot antibody design in a 24-well plate. *bioRxiv*, 2025. Preprint.

[45] N. Skar-Gislinge et al. Protein quaternary structures in solution are a mixture of multiple forms. *Chemical Science*, 13:13058–13069, 2022.

[46] Jian Li and Jianlin Cheng. Enhancing alphafold-multimer-based protein complex structure prediction with multicom in casp15. *Communications Biology*, 6(1):1–14, 2023.

[47] H. Schweke et al. An atlas of protein homo-oligomerization across domains of life. *Cell*, 2024.

[48] Kevin Drew, John B. Wallingford, and Edward M. Marcotte. hu.map 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Molecular Systems Biology*, 17(5):e10016, May 2021.

A Cosmohedra predictions data table

Table 1: Cosmohedra Protein Complex Prediction Results

PDB Code	**Symmetry Class**	**Cosmohedra Prediction RMSD (Å)**	**PDB Reference Data Resolution (Å)**	**Build Time (s)**
7B3Y	I	4.45	3.7	155.1
1T0T	I	2.71	1.75	170.34
4V4M	I	4.56	1.45	203.96
4RFT	I	4.97	3.1	173.72
2WQT	I	3.12	2.8	163
1HQK	I	3.16	1.6	166.3
1JH5	I	6.46	3	102.84
6ZLO	I	1.11	2.9	352
1X36	I	3.03	2.7	68
6HSB	T	3.7	1.95	77.59
2CC9	T	1.29	1.55	97.28
4DCL	T	0.98	3.35	141.5
3N1I	T	2.01	2.2	87.1
2QQY	T	2.99	2	111.96
4CY9	T	1.41	1.78	89.78
3LEO	T	3.48	2.1	78.24
2ZY2	T	4.35	3.3	397.44
6M8V	T	2.71	2.22	136.7
7Q03	T	1.51	2.1	183
6H05	O	4.63	2.9	127.2
7PF9	O	4.44	1.55	177.87
2XZZ	O	1.86	2.3	41.6
3BXV	O	4.29	2.7	152.5
3WIS	O	2.98	1.9	154.1
5H46	O	3.06	2.85	96.9
7OHF	O	3.49	3	147.4
7O63	O	5.44	1.16	94.4
5EKW	O	1.35	1.1	143.9
9JSN	C	3.88	3.73	-
6U23	C	4.95	2.48	-
8UYO	I	3.9	3.3	-
8RB3	I	3.52	3.2	-
8F4X	I	0.91	3.01	-
8V7X	I	3.03	2.8	-
6TAQ	I	4.13	3.9	-

Table includes predictions by past versions of Cosmohedra, for which runtimes were not recorded. I - Icosahedral. T - Tetrahedral. O - Octahedral. C - Cyclic.