Context-aware geometric deep learning for protein sequence design

Lucien F. Krapp, Fernando A. Meireles, Luciano A. Abriata, Matteo Dal Peraro*

Institute of Bioengineering, School of Life Sciences, Ecole Fédérale de Lausanne (EPFL) and Swiss Institute of Bioinformatics (SIB), Lausanne 1015, Switzerland

* To whom correspondence should be addressed: M.D.P., email: matteo.dalperaro@epfl.ch

Protein design and engineering are evolving at an unprecedented pace leveraging the advances of deep learning. Current models nonetheless cannot natively consider non-protein entities within the design process. Here we introduce a deep learning approach based solely on a geometric transformer of atomic coordinates that predicts protein sequences from backbone scaffolds aware of the restraints imposed by diverse molecular environments. This new concept is anticipated to improve the design versatility for engineering proteins with desired functions.

Designing proteins *de novo* to engineer their properties for functional tasks is a grand challenge with direct implications for biology, medicine, biotechnology, and materials science. While physics-based approaches have had success in finding amino acid sequences that fold to a given protein structure, deep learning methods have recently brought a dramatic acceleration by enhancing the design success rates and versatility. Among the most recent and notable examples, ProteinMPNN, based on an encoder-decoder neural network, is able to generate protein sequences experimentally proven to fold as intended[1,2]. More recently, coupled with denoising diffusion probabilistic models for the generation of protein backbones, ProteinMPNN in RFdiffusion has shown remarkable success.[3] In addition, ESM-IF1, based on a protein language model, is capable of generating highly diverse proteins well outside the known universe of natural sequences[4,5]. The model has also recently found experimental validation reporting a very high success rate[6]. Deep learning approaches are however pervasive in the field finding broad application in several protein design tasks[7–10], like for example MaSIF which specializes in the design of protein interactions via learned protein surface fingerprints[11,12].

Although these models can natively handle multiple protein chains in their inputs, and as such they can design the sequences of interacting proteins, they cannot natively consider non-protein entities within the design process, which hampers their versatility and limit their spectrum of application. Here, to address this limitation, we introduce CARBonAra (namely, Context-aware Amino acid Recovery from Backbone Atoms and heteroatoms), a new protein sequence generator model based on our recent Protein Structure Transformer (PeSTo[13]), a geometric transformer architecture that operates on atom point clouds. Representing molecules uniquely by element names and coordinates, PeSTo's transformer can be applied to and predict protein interfaces with virtually any kind of molecules, either other proteins, nucleic acids,

lipids, ions, small ligands, or cofactors. Based on the same architecture, trained uniquely on structural data available on the PDB, CARBonAra predicts the amino acid confidence per position from a backbone scaffold alone or complexed by any kind of non-protein molecules. The model uses geometrical transformers to encode the local neighbourhood of the atomic point cloud using the geometry and atomic elements. It encodes the interactions of the nearest neighbours and employs a transformer to decode and update the state of each atom. By pooling the atom states from the atomic to the residue level and decoding them, the model predicts multi-class residue-wise amino acid confidences (**Figure 1a** and **Methods**). CARBonAra thus provides a potential sequence space that can be refined through the incorporation of specific constraints, such as a molecular context critical to the protein's function, a particular objective, or provided allowed conformations. CARBonAra offers a novel level of flexibility in protein design by recognizing and incorporating any molecular context into its sequence predictions. This distinctive capability of our method expands therefore the scope of applications in the field of protein design.

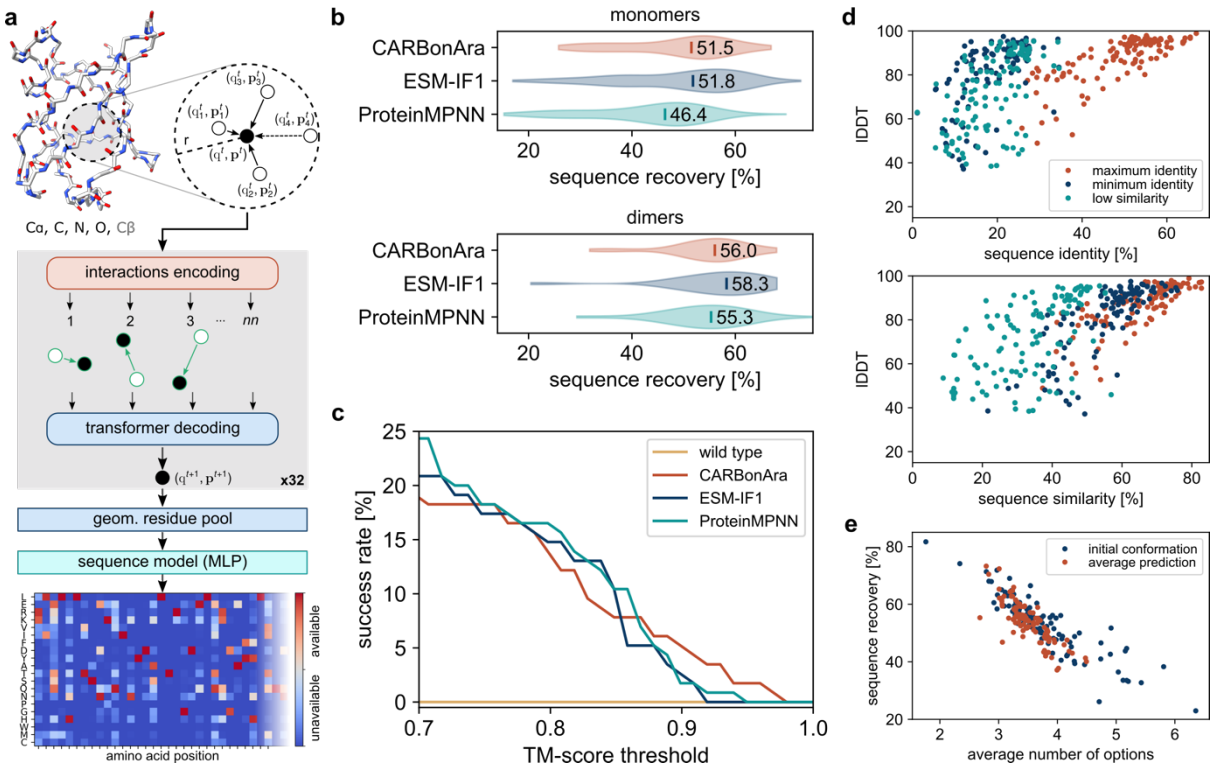


Figure 1 | CARBonAra architecture and comparison with other state-of-the-art methods. (**a**) The model applies multiple geometric transformer operations to the coordinates and atom element of a backbone scaffold with added virtual C_β to predict the amino acid confidence at each position in the sequence. (**b**) Comparison of the sequence recovery of different methods for monomers and dimers with indicated median sequence recovery. (**c**) Percentage of AlphaFold predicted structures, in single sequence mode, above a TM-score threshold. (**d**) Local Distance Difference Test (lDDT) of AlphaFold predicted structures against scaffold monomers from sequences generated using CARBonAra with, as objective, maximum sequence identity, minimum sequence identity, and low sequence similarity. (**e**) Comparison of the sequence recovery between the predicted sequence on crystal structures and the consensus sequence predictions derived from 500 frames sampled from 1 µs molecular dynamics simulations for 80 monomers.

CARBonAra performs on par with state-of-the-art methods like ProteinMPNN and ESM-IF1 for sequence prediction of isolated proteins or protein complexes (**Figure 1b**), while having a similar computational cost taking only a few seconds per run (~3 seconds). Our method achieves a median sequence recovery rate of 51.3% for protein monomer design and 56.0% for dimer design when reconstructing protein sequences from backbone structures. Moreover, the success rate of the generated sequences using AlphaFold in single-sequence mode is commendable, especially in generating structures with a TM-score above 0.9 (**Figure 1c**). We observed that the model is able to learn the tighter amino acid packing at protein cores thus resulting in higher recovery rates and fewer amino acid possibilities for buried amino acids (**Supplementary Figure 1a-c**). As such, CARBonAra confidently recovers core amino acids while demonstrating greater flexibility on the protein's surface, unless additional functional or structural constraints are provided.

In contrast to other methods, CARBonAra uses multi-class amino acid predictions that generate a space of potential sequences, opening various possibilities for sequence sampling. For example, one can tailor sequences to meet specific objectives, such as achieving maximal or minimal sequence identity, or low sequence similarity in order to design unique sequences with a specific fold (**Figure 1d** and **Supplementary Figure 2,** see also Methods). An informative way to refine the sequence space uses dynamics as a constraint. By applying CARBonAra to structural trajectories from molecular dynamics (MD) simulations, we were able to improve sequence recovery, especially in cases that previously showed low recovery rates (**Figure 1e**). Simultaneously, we observed a reduction in the number of possible amino acids predicted per position. This further limit the sequence space and could enable the design of targeted structural conformations.

More importantly, leveraging PeSTo's architecture, this model has the new ability to perform protein sequence prediction conditioned by a specific non-protein molecular context. On a test set similar to the one used for PeSTo, we show that the overall structure median sequence recovery increased from 54% to 58% (**Supplementary Figure 3**) when an additional molecular context is provided. In particular, CARBonAra achieves median sequence recovery rates at the interface of 56% when protein interacting partners are considered and 55% when nucleic acids are used as interfacial restraints, providing a significant improvement over predictions without context (**Figure 2a**). Similarly, the recovery rate at the interface improved significantly if small-molecule entities such as ions (67%), lipids (57%), and ligands (60%) are included (**Figure 2a**). Including these molecules boosts sequence recovery in their surroundings, and reduces the number of amino acid possibilities to sample from (**Figure 2b**). This shows CARBonAra's power to properly craft the residue types required for the binding of specific molecules.

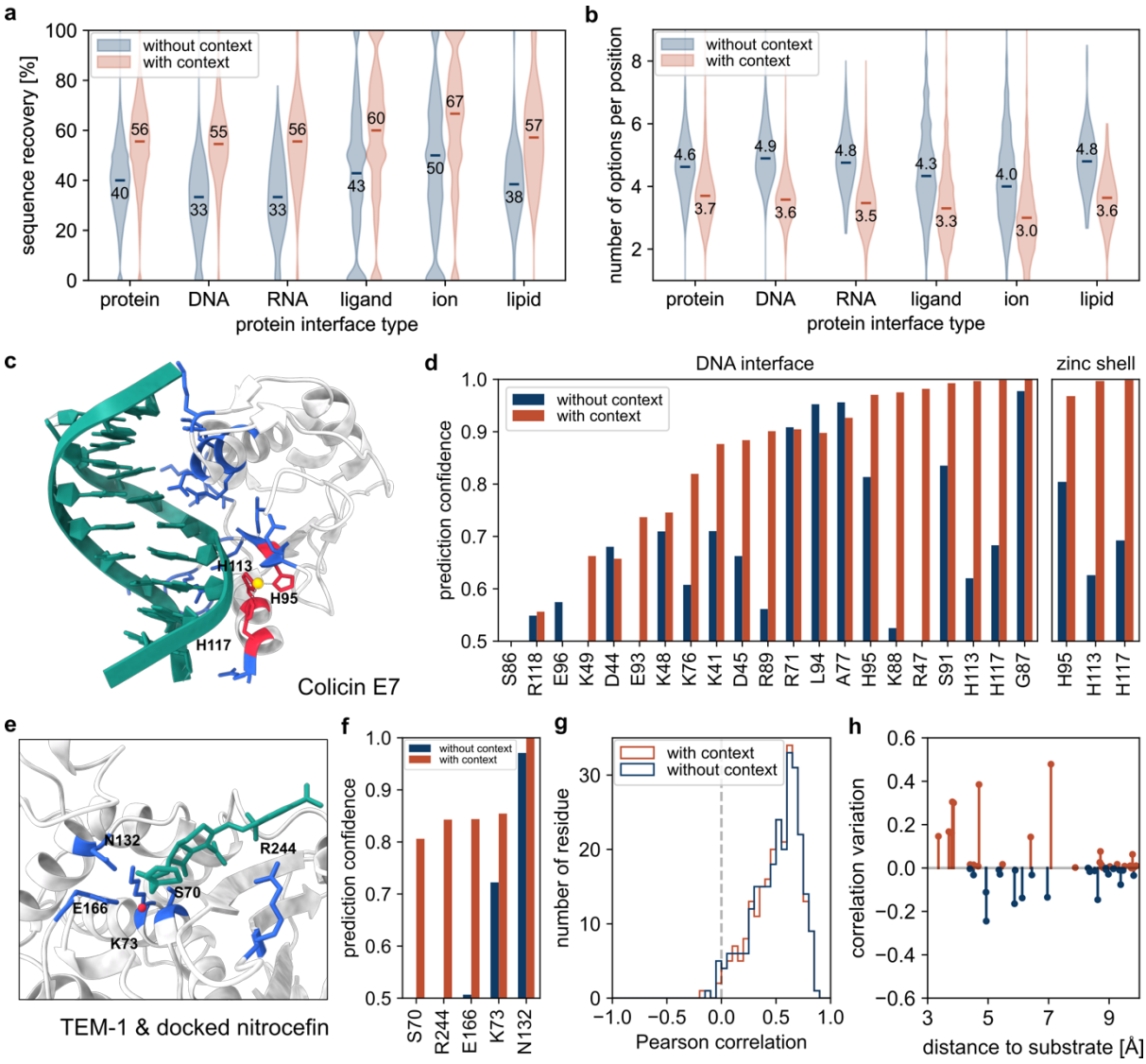


Figure 2 | Context-aware amino acid recovery extends to various biomolecules. (**a**) Sequence recovery at the interface (residues within 5 Å) without and with proteins, nucleic acids, ligands, ions, and lipids binders. (**b**) Number of predicted possible amino acids per position at the interface (residues within 5 Å) without and with proteins, nucleic acids, ligands, ions, and lipids binders (considering a confidence prediction threshold of 0.5). (**c**) Colicin E7 endonuclease domain in complex with DNA and a zinc ion (PDB: 1ZNS). The protein-DNA interface (residues within 4 Å) is highlighted in blue. The protein-zinc shell is highlighted in red (residues within 3 Å). (**d**) Estimated accurate prediction probability for the scaffold amino acids at the protein-DNA interface and the protein-zinc shell with and without the presence of DNA and zinc. (**e**) Nitrocefin docked in the active site of the β-lactamase TEM-1 (PDB: 1BT5). Relevant residues for substrate recognition and hydrolysis are shown in blue, nitrocefin in green, and the catalytic water molecule in red. (**f**) Prediction confidence with and without the substrate for the relevant amino acids for binding. (**g**) Correlation of the predictions with deep sequencing analysis of TEM-1. (**h**) Correlation variation by adding the context (nitrocefin and catalytic water) for the amino acids close (in C_β distance) to the substrate.

An exemplary case to illustrate the power of this approach is the endonuclease domain of ColE7, which interacts with duplex DNA in a zinc-dependent manner[14]. The sequence recovery rate obtained by CARBonAra showed a significant increase from 29% to 52% at the metal and DNA interfaces when the zinc ion or the 12-bp DNA duplex was included as resolved in the native structure (**Figure 2d**). Thus, imposing

the presence of non-protein interacting interfaces can enhance the sequence recovery rate significantly, also with respect to predictions done by ProteinMPNN (24%) and ESM-IF1 (43%) (**Supplementary Table 1**). Interestingly, when a non-native molecular context is provided such as a larger ion (e.g., calcium) the sequence recovery rate decreased (**Supplementary Figure 4**). Thus, the predicted amino acid confidence of an ion pocket is widely dependent on the given context, as illustrated also for the metallo β-lactamase BJP-1 (**Supplementary Figure 5**).

Relevant for enzyme design is the possibility to design sequences under the restraints provided by a desired substrate or high-affinity ligand. To test this case, we explored CARBonAra's ability to predict the sequence of a TEM-1 β-lactamase-like enzyme when the native context at the active site is provided (**Figure 2e**). Without context, the catalytic S70 and substrate binding R244 are never predicted positively (confidence of 0.39 and 0.11 respectively, **Figure 2f**), however, when the prediction is done with a β-lactam (here nitrocefin) docked at the catalytic pocket, the catalytic triad S70, K73, and E166, along with key residues necessary to β-lactam binding (i.e., N132, R244) all have a high prediction confidence (> 0.8) and low ranking (top 2) (**Supplementary Figure 6**). Importantly, in this case, the sequence recovery is maximal when also the catalytic water is considered, hinting at a very high sensitivity for the molecular context.

Given that TEM-1 β-lactamase has been widely studied, we took the occasion to probe what information CARBonAra's residue-wise amino acid probabilities provide when compared to experimental data. We correlated the estimated probabilities to the residue-wise amino acid probabilities measured experimentally through deep sequencing of a saturated mutagenesis library of the TEM-1 β-lactamase[15] (**Figure 2g**). We observed an average correlation of 0.51 ± 0.21 for CARBonARa with deep sequencing data, which is similar to the correlation between the deep sequencing data with the multiple sequence alignment of this enzyme's family (0.52 ± 0.22). This shows that CARBonARa's estimated probabilities can capture functional sequence variability, a central topic in the realm of protein evolution[16,17]. Moreover, we observed that adding the context to the active site of TEM-1 (i.e. docked nitrocefin and the catalytic water) improved the correlation locally (i.e. for amino acids within 5 Å) but also affects the predictions of amino acids further away (up to 10 Å). These results hint at the possibility to use CARBonAra for the study of the effect of a specific context locally as well as their long-range influences (**Figure 2h**).

In summary, CARBonAra is a novel, structure-centric way to predict protein sequences given a backbone geometry. Leveraging a geometric transformer architecture that can handle any molecular context, CARBonAra can optimize the sequence prediction by considering not only interacting protein partners but also nucleic acids, lipids, small molecules, and ions. This new concept, coupled with the possibility to sample backbone conformations using modern diffusion models, opens exciting new possibilities for the functional design of novel protein-based materials and therapeutics.

Availability. All datasets, methods and the CARBonAra source code presented in this work will be released upon publication at https://github.com/LBM-EPFL/CARBonARa.

Acknowledgments. The Swiss National Science Foundation is acknowledged for supporting this work (grant number 205321_192371 to MDP). We also acknowledge the Swiss National Supercomputing Centre (CSCS) for the generous computing time allocation used to run molecular dynamics simulations.

Author contributions. LFK and MDP conceived and designed the research project. LFK designed and implemented the CARBonAra code. LFK, FAM, LAA, and MDP analysed the data. LFK, FAM, LAA, and MDP wrote the paper.

References

1. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
2. Wicky, B. I. M. *et al.* Hallucinating symmetric protein assemblies. *Science* **378**, 56–61 (2022).
3. Watson, J. L. *et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models*. http://biorxiv.org/lookup/doi/10.1101/2022.12.09.519842 (2022) doi:10.1101/2022.12.09.519842.
4. Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. 2022.04.10.487779 Preprint at https://doi.org/10.1101/2022.04.10.487779 (2022).
5. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
6. Verkuil, R. *et al.* Language models generalize beyond natural proteins. 2022.12.21.521521 Preprint at https://doi.org/10.1101/2022.12.21.521521 (2022).
7. Ingraham, J., Garg, V., Barzilay, R. & Jaakkola, T. Generative Models for Graph-Based Protein Design. in *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).
8. Sgarbossa, D., Lupo, U. & Bitbol, A.-F. Generative power of a protein language model trained on multiple sequence alignments. *eLife* **12**, e79854 (2023).
9. Zhou, X. *et al. Protein Sequence Design by Entropy-based Iterative Refinement*. http://biorxiv.org/lookup/doi/10.1101/2023.02.04.527099 (2023) doi:10.1101/2023.02.04.527099.
10. Lisanza, S. L. *et al.* Joint Generation of Protein Sequence and Structure with RoseTTAFold Sequence Space Diffusion.
11. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* **17**, 184–192 (2020).

12. Gainza, P. *et al.* De novo design of protein interactions with learned surface fingerprints. *Nature* **617**, 176–184 (2023).
13. Krapp, L. F., Abriata, L. A., Cortés Rodriguez, F. & Dal Peraro, M. PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nat Commun* **14**, 2175 (2023).
14. Doudeva, L. G. *et al.* Crystal structural analysis and metal-dependent stability and activity studies of the ColE7 endonuclease domain in complex with DNA/Zn2+ or inhibitor/Ni2+. *Protein Science* **15**, 269–280 (2006).
15. Deng, Z. *et al.* Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. *J Mol Biol* **424**, 150–167 (2012).
16. Abriata, L. A., Palzkill, T. & Dal Peraro, M. How structural and physicochemical determinants shape sequence constraints in a functional enzyme. *PLoS One* **10**, e0118684 (2015).
17. Mayorov, A., Dal Peraro, M. & Abriata, L. A. Active Site-Induced Evolutionary Constraints Follow Fold Polarity Principles in Soluble Globular Enzymes. *Mol Biol Evol* **36**, 1728–1733 (2019).
18. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915–10919 (1992).
19. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
20. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).
21. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
22. Evans, R. *et al. Protein complex prediction with AlphaFold-Multimer*. http://biorxiv.org/lookup/doi/10.1101/2021.10.04.463034 (2021) doi:10.1101/2021.10.04.463034.
23. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **57**, 702–710 (2004).
24. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).
25. Vreven, T. *et al.* Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *Journal of Molecular Biology* **427**, 3031–3041 (2015).
26. Abriata, L. A. & Dal Peraro, M. Assessment of transferable forcefields for protein simulations attests improved description of disordered states and secondary structure propensities, and hints at multi-protein systems as the next challenge for optimization. *Computational and Structural Biotechnology Journal* **19**, 2626–2636 (2021).
27. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* **14**, 71–73 (2017).

28. Van Der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry* **26**, 1701–1718 (2005).
29. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J Comput Chem* **31**, 455–461 (2010).

Supplementary Information

Algorithm 1: Geometric transformer

Input: Center node features: $q \in \mathbb{R}^{N \times S}, p \in \mathbb{R}^{N \times S \times 3}$
Context neighbors features: $q_{nn} \in \mathbb{R}^{N \times n \times S}, p_{nn} \in \mathbb{R}^{N \times n \times S \times 3}$
Geometry features: $d_{nn} \in \mathbb{R}^{N \times n}, r_{nn} \in \mathbb{R}^{N \times n \times 3}$
Output: New state of center node: $q', \vec{p}'$

// pack node and edges features
$X_n \leftarrow concat(q, \|\vec{p}\|) \in \mathbb{R}^{N \times 2S}$ ▷ Node features
$X_e \leftarrow concat(d_{nn}, q, \|\vec{p}\|, q_{nn}, \|\vec{p}_{nn}\|, \vec{p} \cdot \vec{r}_{nn}, \vec{p}_{nn} \cdot \vec{r}_{nn}) \in \mathbb{R}^{N \times n \times 6S+1}$ ▷ Edges features

// encode queries from node state
$Q_q, Q_p \leftarrow f_{nqm}(X_n) \in \mathbb{R}^{N \times N_h \times N_k} \times \mathbb{R}^{N \times N_h \times N_k}$ ▷ Encoded queries

// encode keys from edges state
$K_q \leftarrow f_{eqkm}(X_e) \in \mathbb{R}^{N \times n \times N_k}$ ▷ Scalar keys
$K_p \leftarrow f_{epkm}(X_p) \in \mathbb{R}^{N \times 3n \times N_k}$ ▷ Vector keys

// encode values from edges state
$V_q, V_p \leftarrow f_{evm}(X_e) \in \mathbb{R}^{N \times n \times S} \times \mathbb{R}^{N \times n \times S}$ ▷ Edges encoded values
$\vec{X}_g \leftarrow concat(V_p \cdot \vec{r}_{nn}, \vec{p}, \vec{p}_{nn}) \in \mathbb{R}^{N \times 3n \times S \times 3}$ ▷ Geometric features

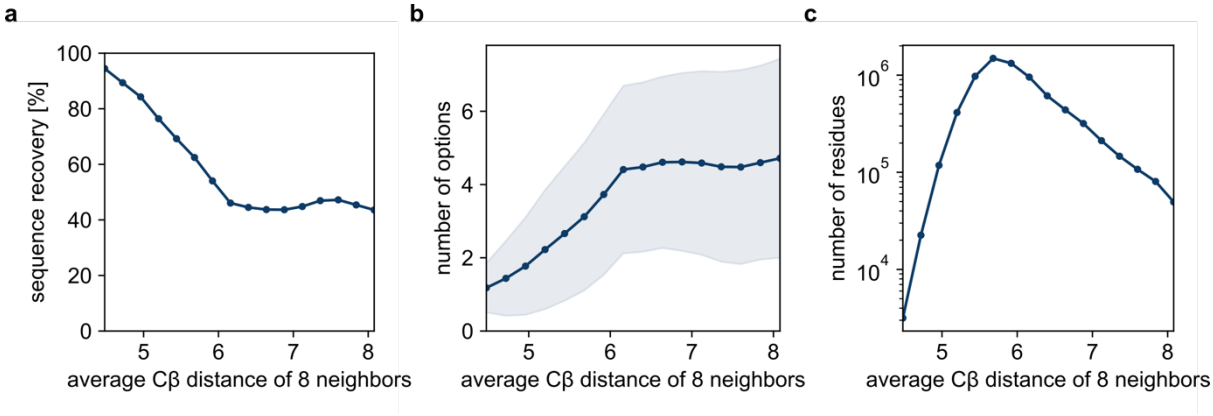
// scaled dot-product attention and projection
$q_h \leftarrow f_{qpm}(Attention(Q_q, K_q, V_q)) \in \mathbb{R}^{N \times S}$ ▷ Scalar hidden state
$\vec{p}_h \leftarrow f_{ppm}(Attention(Q_p, K_p, \vec{X}_g)) \in \mathbb{R}^{N \times S \times 3}$ ▷ Vectorial hidden state

// update state with residual
$q' \leftarrow q + q_h$
$\vec{p}' \leftarrow \vec{p} + \vec{p}_h$

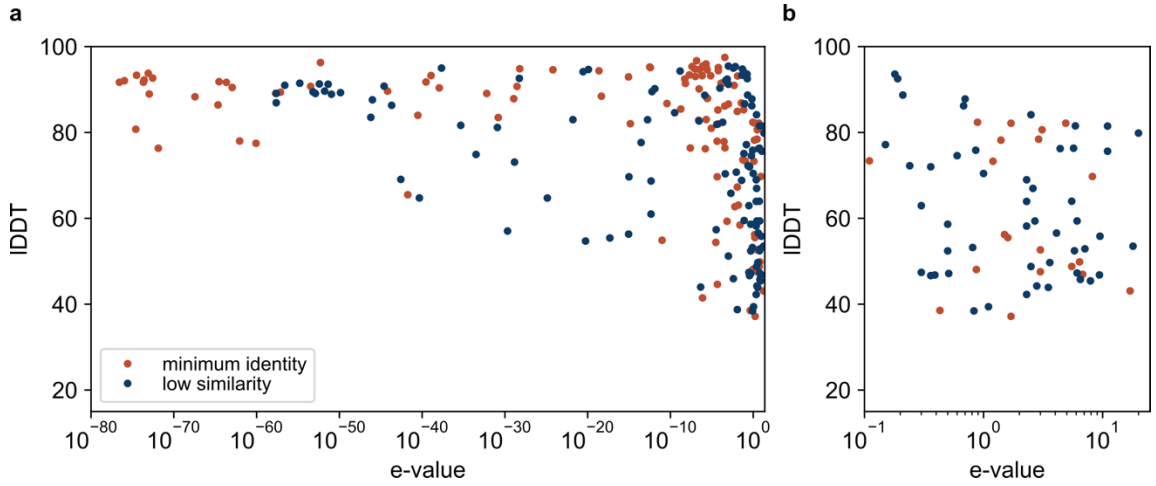
Supplementary Algorithm 1 | CARBonAra geometric transformer. Based on the PeSTo architecture, each geometric transformer is composed of 5 neural networks of 3 layers with an exponential linear unit (ELU) activation function. The characteristic dimensions are the number of atoms (N), the state size (S), the number of nearest neighbors (nn), the dimension of the embedding for the keys (N_k) and the number of attention heads (Nh). The neural networks haven a flat architecture with hidden layers width equal to the input and output state size (S). The multi-layers perceptrons (MLP) are the node query model (f_{nqm}), encoding scalar key model (f_{eqkm}), encoding vector key model (f_{epkm}), encoding value model (f_{evm}), and scalar state projection model (f_{qpm}). The vectorial hidden state is projected over the attention heads with a weighted sum (W_{ppm}) to preserve the rotation equivariance of the operation. The output vector state belongs to the span of the geometry and vector states.

Supplementary Table 1 | Method comparison with and without DNA bound on Colicin E7. Sequence recovery and similarity for the whole structure and at the interface (residues within 4 Å) with and without DNA of Colicin E7.

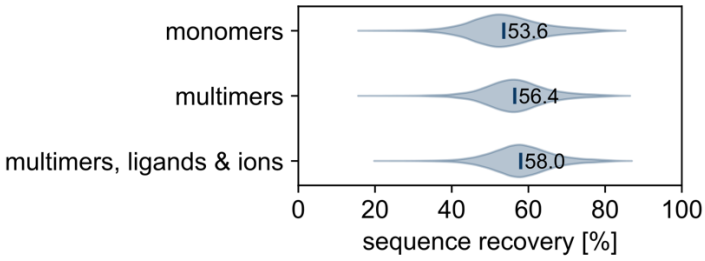
	sequence recovery	sequence similarity	interface sequence recovery	interface sequence similarity
CARBonAra without context	47.5%	71.2%	28.6%	52.4%
CARBonAra with context	49.2%	**72.0%**	**52.4%**	**71.4%**
ProteinMPNN	39.8%	61.0%	23.8%	61.9%
ESM-IF1	**50.8%**	66.9%	42.9%	**71.4%**



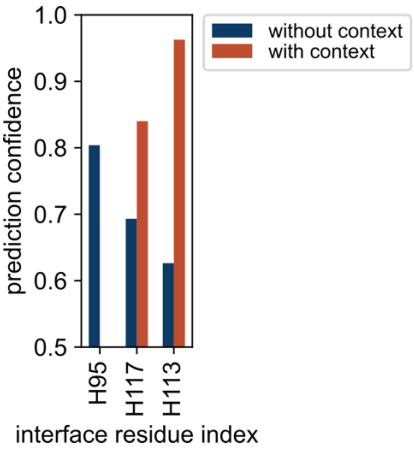
Supplementary Figure 1 | Analysis of buried against surface amino acids. (**a**) Sequence recovery, (**b**) number of predicted options per position and (**c**) number of residues as a function of the average C_β distance of the 8 nearest neighbours (18866 structures from the testing dataset).



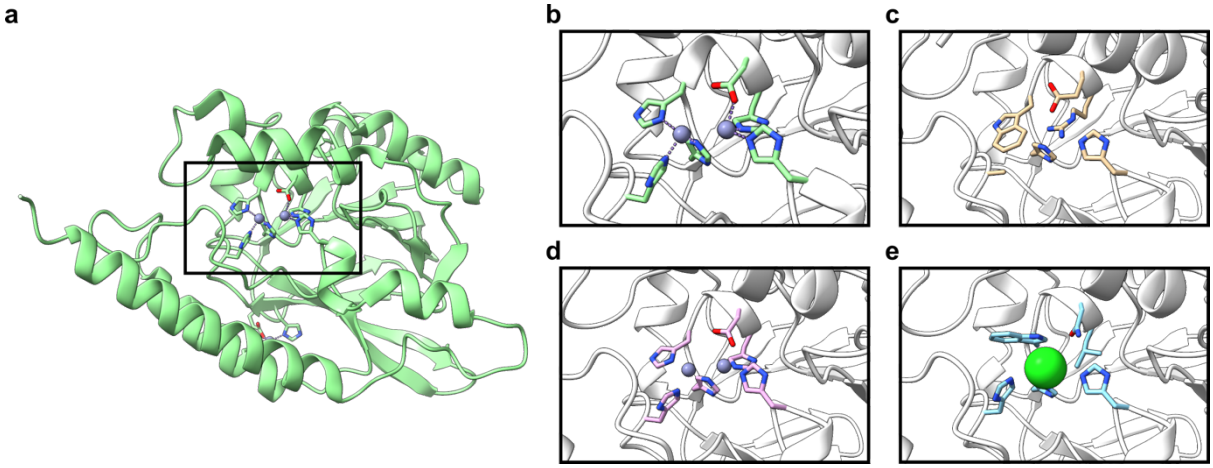
Supplementary Figure 2 | Analysis of the generated sequences. (**a**) lDDT of the AlphaFold predicted structures as a function of the expect value (E-value) of the generated sequences. (**b**) Close up on the generated sequences with a high E-value.



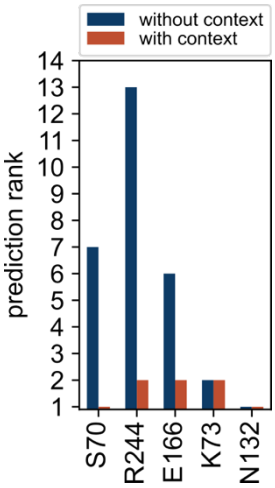
Supplementary Figure 3 | Benchmark of different use cases. Sequence recovery distribution for systems of monomers, multimers and any biomolecules (18866 structures from the testing dataset). The median sequence recovery is indicated for each case.



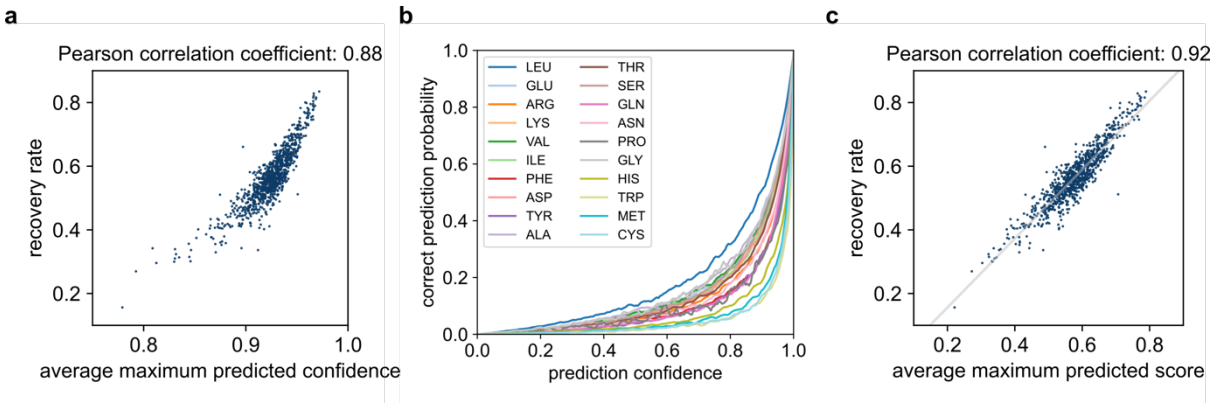
Supplementary Figure 4 | Effect of changing the ion type on the prediction. The prediction confidence for the three most important amino acids for ion binding in the case where the zinc ion of Colicin E7 is replaced with a calcium ion.



Supplementary Figure 5 | Effect of the ion context on the optimal predicted sequence in the case of a metallo β-lactamase zinc binding pocket. (**a**) Metallo β-lactamase structure with a pocket containing two zinc ions (PDB ID: 3LVZ). (**b**) WT pocket of the metallo β-lactamase. Pocket of an AlphaFold predicted structure with a designed sequence applied to the scaffold structure without zinc ions (**c**), containing the original zinc ions (**d**) and containing a manually placed chloride ion (**e**).



Supplementary Figure 6 | Effect of the docked nitrocefin and catalytic water in TEM-1 on the prediction ranking. Rank of the prediction from maximum to minimum confidence for the 5 important amino acids at the pocket without and with the docked nitrocefin and catalytic water.



Supplementary Figure 7 | Prediction confidence analysis. (**a**) Recovery rate as a function of the average maximum prediction score (943 structures from the testing dataset). (**b**) Relationship between prediction confidence and the prediction accuracy for each amino acid type (4096 subunits from the training dataset). (**c**) Rescaling prediction score into a prediction confidence correlated with the probability to be correct (943 structures from the testing dataset).