Context-aware geometric deep learning for protein sequence design

Lucien F. Krapp[1,2], Fernando A. Meireles[1,2], Luciano A. Abriata[1,2], Jean Devillard[1], Sarah Vacle[1,2], Maria J. Marcaida[1,2], Matteo Dal Peraro[1,2,*]

[1] Laboratory for Biomolecular Modeling, Institute of Bioengineering, School of Life Sciences, Ecole Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland

[2] Swiss Institute of Bioinformatics (SIB), Lausanne 1015, Switzerland

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

Supplementary Information

- Supplementary Algorithm 1;
- Supplementary Figures 1-14;
- Supplementary Tables 1-2;
- Supplementary Dataset 1;

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

$K_p \leftarrow f_{epkm}(X_e) \in \mathbb{R}^{N \times 3n \times N_k}$ ▷ 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

$\vec{X}_g \leftarrow \text{concat}(V_p \odot \vec{r}_{nn}, \vec{p}, \vec{p}_{nn}) \in \mathbb{R}^{N \times 3n \times S \times 3}$ ▷ Geometric features

`// scaled dot-product attention and projection`

$q_h \leftarrow f_{qpm}(\text{softmax}(\frac{Q_q K_q^T}{\sqrt{N_k}})V_q) \in \mathbb{R}^{N \times S}$ ▷ Scalar hidden state

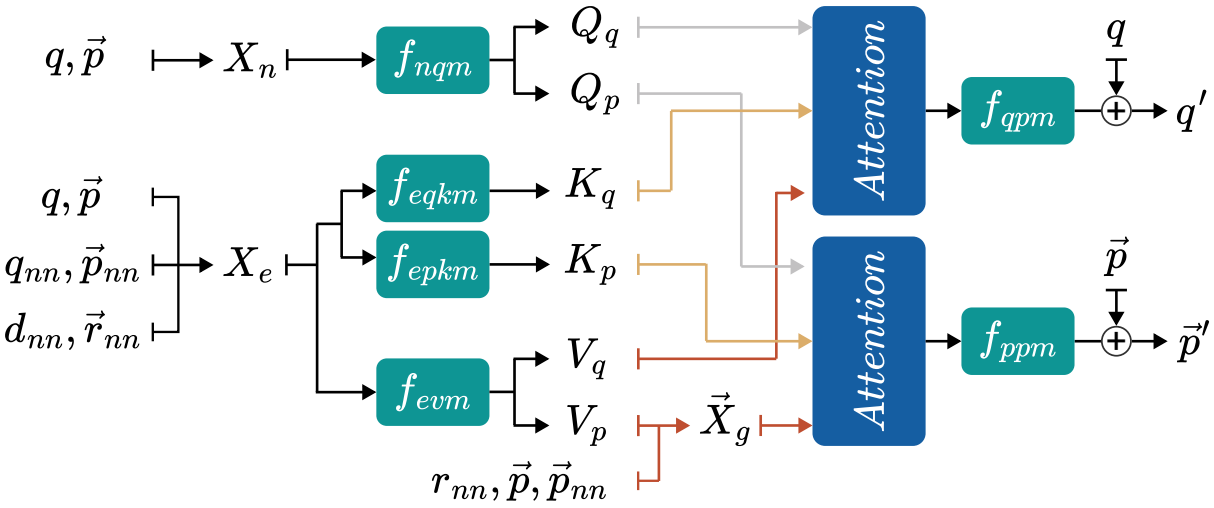
$\vec{p}_h \leftarrow W_{ppm}\ \text{softmax}(\frac{Q_p K_p^T}{\sqrt{N_k}})\vec{X}_g \in \mathbb{R}^{N \times S \times 3}$ ▷ Vectorial hidden state

`// update state with residual`

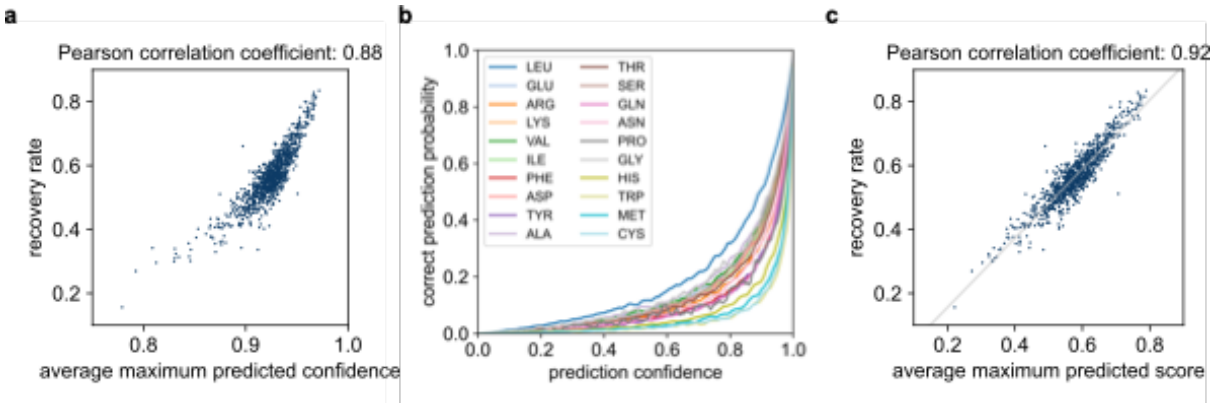
$q' \leftarrow q + q_h$

$\vec{p}' \leftarrow \vec{p} + \vec{p}_h$

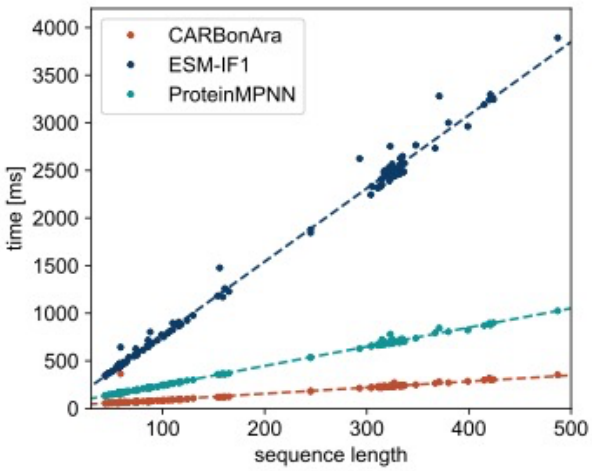
Supplementary Algorithm 1 | CARBonAra geometric transformer. Based on the PeSTo architecture, each geometric transformer is composed of 5 neural networks of 3 layers with an exponential linear unit (ELU) activation function. The characteristic dimensions are the number of atoms (N), the state size (S), the number of nearest neighbors (nn), the dimension of the embedding for the keys (N_k) and the number of attention heads (N_h). 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.



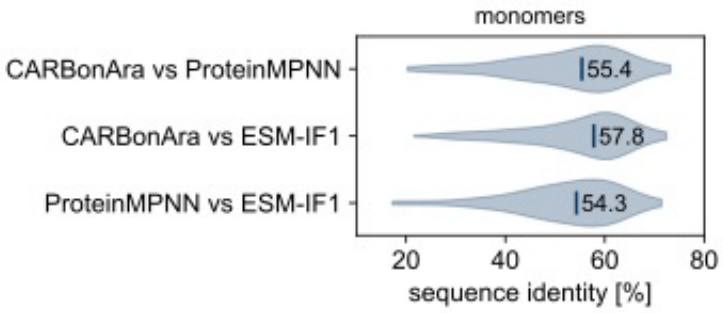
Supplementary Figure 1 | CARBonAra geometric transformer. The inputs of the geometric transformer are scalar state (q) and vector state (**p**) for the central atom and the neighbors scalar states (q_{nn}), vector states ($\mathbf{p}_{nn}$), distances (d_{nn}) and relative displacement vectors ($\mathbf{r}_{nn}$). First, we extract the scalar information of the central node features (X_n) and edges features (X_e) from the inputs. The central node features produce the queries (Q_q, Q_p) through an MLP (f_{nqm}). The edge node features produce the keys (K_q, K_p) and values (V_q, V_p) through multiple MLP (f_{eqkm}, f_{epkm}, f_{evm}). We project the vector track values (V_p) on relative displacement vectors (r_{nn}) and concatenate the vector states to create the geometric features ($\mathbf{X}_g$). We compute the multi-heads key, query and value attention for the scalar and vector track. We reduce the outputs of the attention operation with an MLP for the scalar quantities (f_{qpm}) and a weighted sum (f_{ppm}) for the vector track to preserve the rotation equivariance of the operation. Lastly, we add the input states as residual connections.



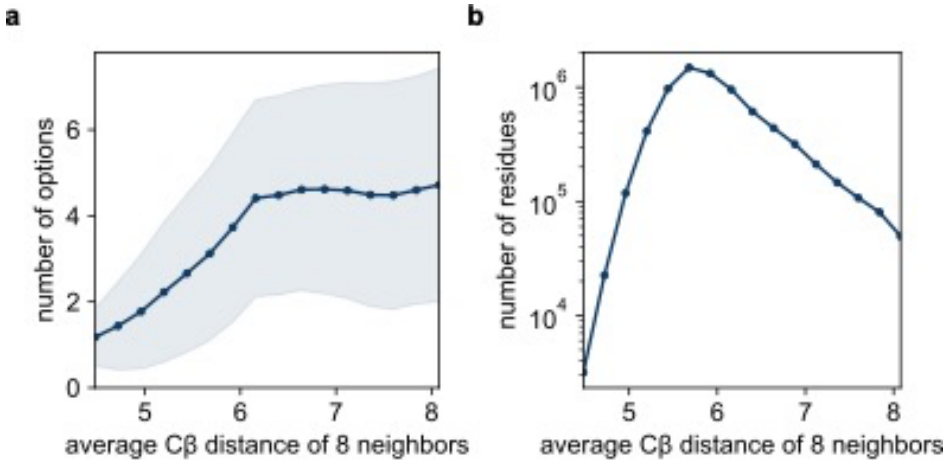
Supplementary Figure 2 | Prediction confidence analysis. (**a**) Recovery rate as a function of the average maximum prediction score (943 structures from the testing dataset). (**b**) Relationship between prediction confidence and the prediction accuracy for each amino acid type (4096 subunits from the training dataset). (**c**) Rescaling prediction score into a prediction confidence correlated with the probability to be correct (943 structures from the testing dataset). This mapping (computed from the training set) converts the prediction confidence into a probability that can be used for sampling.



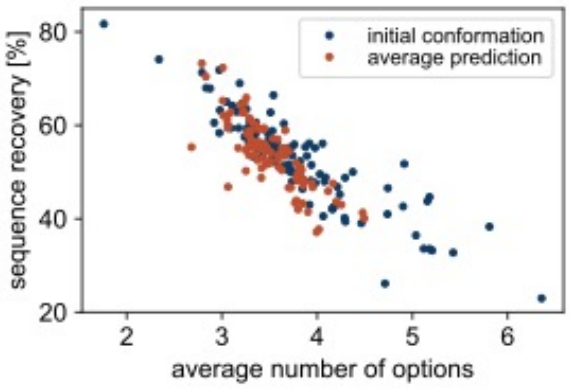
Supplementary Figure 3 | Run time analysis on GPU. Model run time as a function of the sequence length tested on a Nvidia RTX 2080 Ti and Intel i9-9900K. ESM-IF1 runs out of memory on the GPU with larger system so we compared the three methods on 142 structures with sequence length under 500 amino acids.



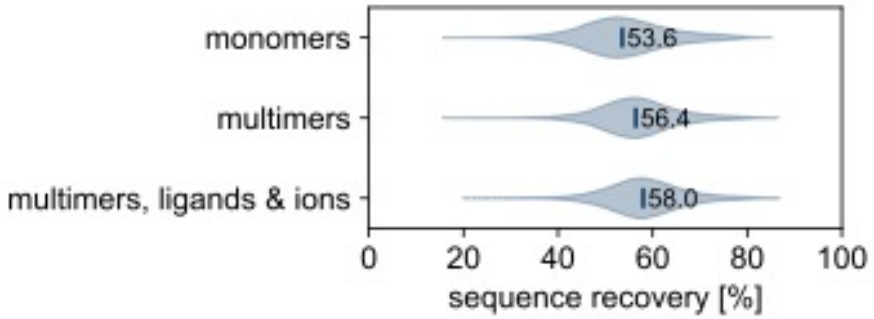
Supplementary Figure 4 | Comparison of predicted sequences between methods. Sequence identity between sequences predicted by two different methods for 142 monomers. The median sequence identity is indicated for the three comparisons. The sequences predicted by the three methods are as similar to each other as to the original scaffold sequence.



Supplementary Figure 5 | Analysis of buried against surface amino acids. (**a**) Number of predicted options per position and (**b**) number of residues as a function of the average C_β distance of the 8 nearest neighbours (18866 structures from the testing dataset).



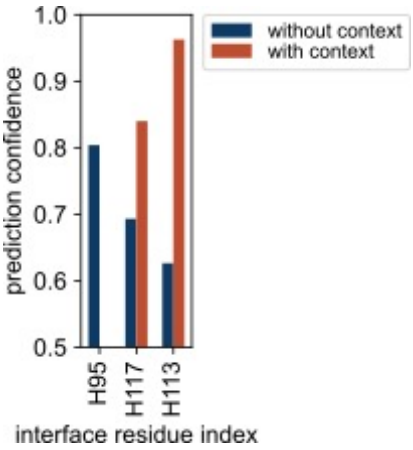
Supplementary Figure 6 | Effect of backbone conformations on predictions. Sequence recovery as a function of the average Cβ distance of the 8 nearest neighbours. Prediction recovery rate against the average number of options for the reference experimental structure (initial conformation) and the consensus sequence predictions (average prediction), derived from 500 frames sampled from 1 μs molecular dynamics simulations for 80 monomers.



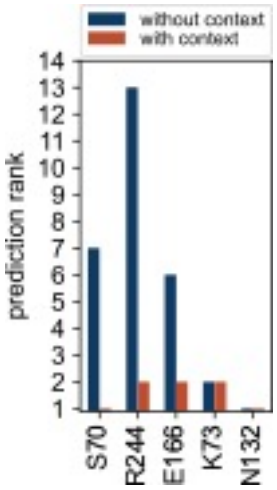
Supplementary Figure 7 | Benchmark of different use cases. Sequence recovery distribution for systems of monomers, multimers and any biomolecules (18866 structures from the testing dataset). The median sequence recovery is indicated for each case.

	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%**

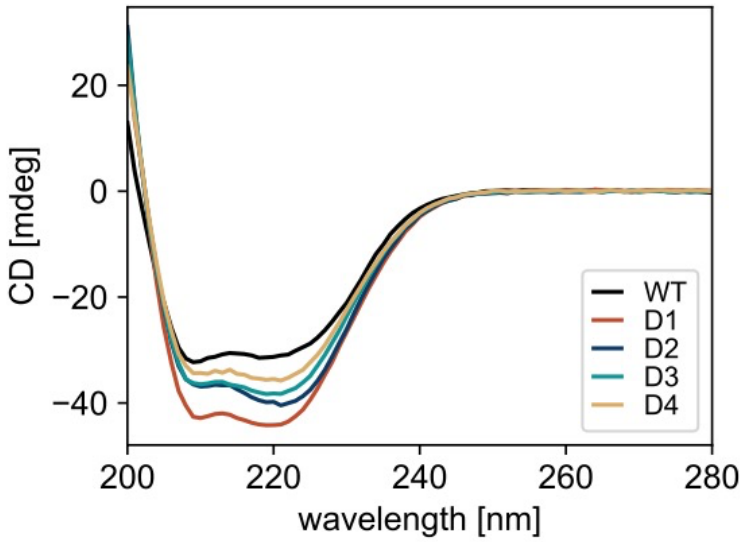
Supplementary Table 1 | Method comparison with and without DNA bound on Colicin E7. Sequence recovery and similarity for the whole structure and at the interface (residues within 4 Å) with and without DNA of Colicin E7.



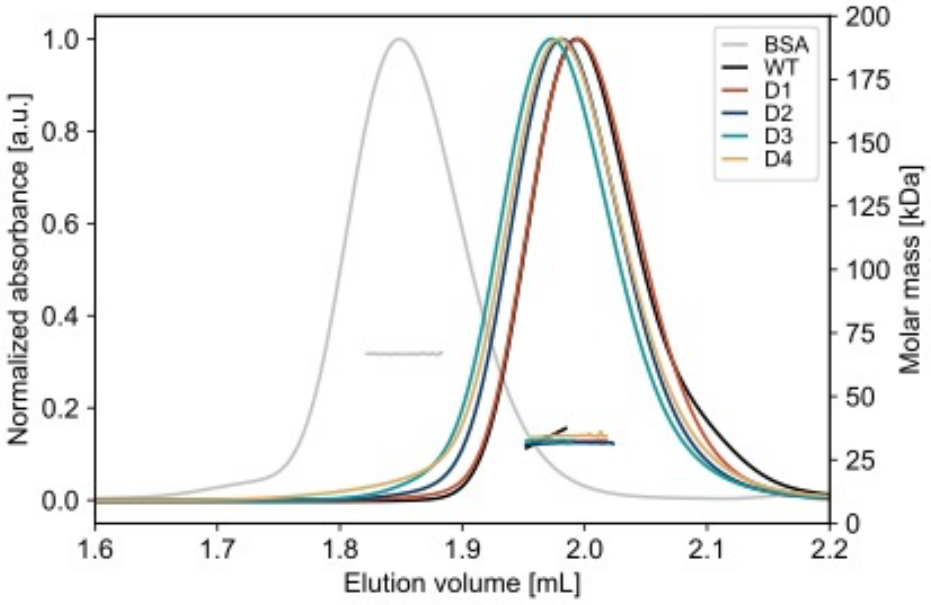
Supplementary Figure 8 | Effect of changing the ion type on the prediction for the case of endonuclease domain of ColE7**.** The prediction confidence for the three most important amino acids for ion binding in the case where the zinc ion of Colicin E7 is replaced with a calcium ion.



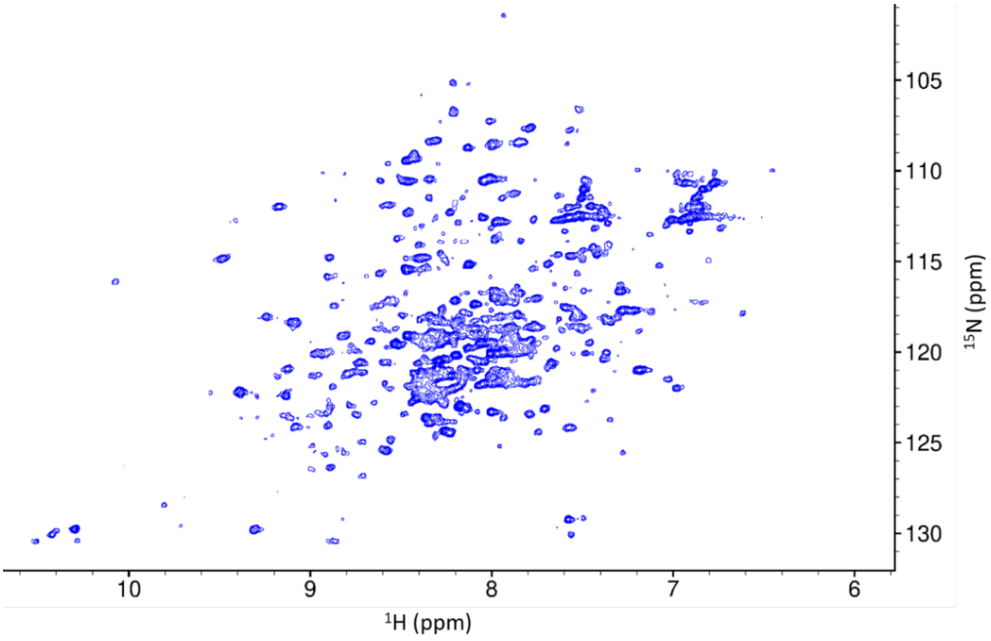
Supplementary Figure 9 | Effect of the docked nitrocefin and catalytic water in TEM-1 on the prediction ranking. Rank of the prediction from maximum to minimum confidence for the 5 important amino acids at the pocket without and with the docked nitrocefin and catalytic water.



Supplementary Figure 10 | TEM-like designs maintain second structural motifs as TEM-1. Far-UV circular dichroism spectra of the 4 soluble designs produced from TEM-1.



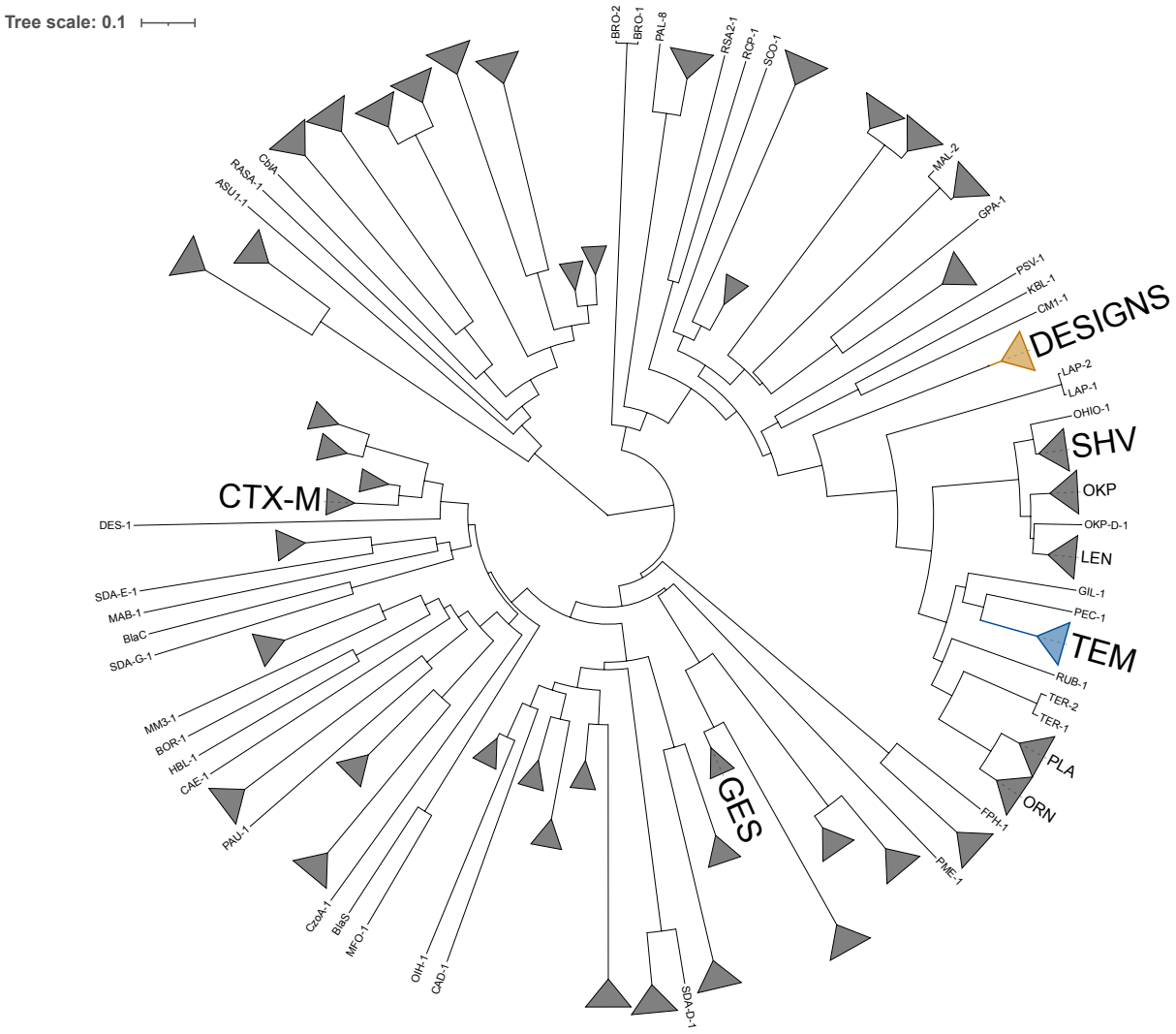
Supplementary Figure 11 | TEM-like designs are monomeric. Size Exclusion Chromatography Multi Angle Light Scattering (SEC-MALS) data show that all the TEM-like designs are monodispersed monomers with molar masses of ~31 kDa, as expected. The 280 nm absorbance trace also shows that they elute at the same elution volume in the Superose 6 column, indicative of similar 3-dimensional structure as the WT TEM-1. BSA (66 kDa) was used as internal calibration of the MALS set up.



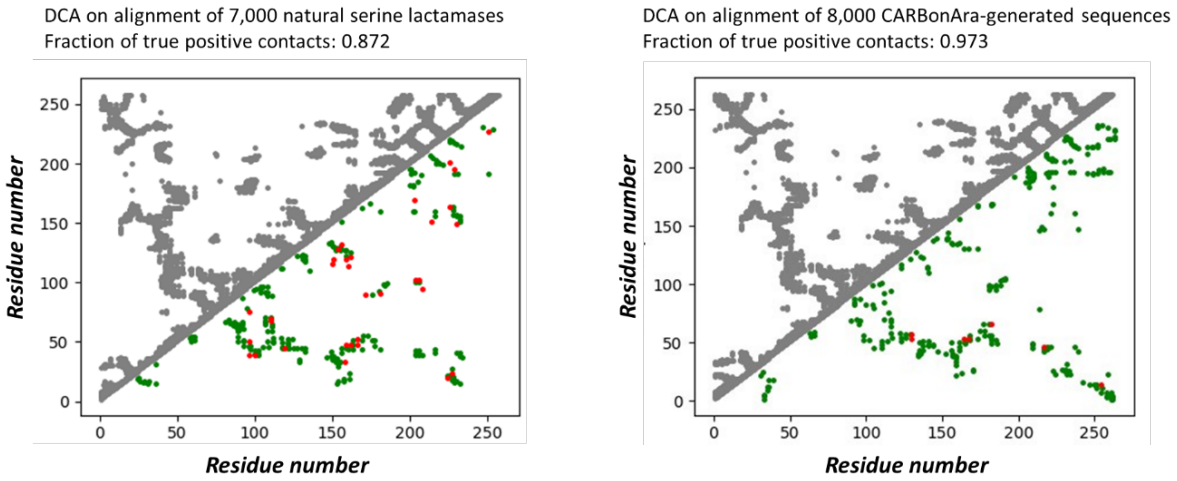
Supplementary Figure 12 | TEM-like design shows typical NMR spectrum of a folded protein. ^{1}H,^{15}N HSQC spectrum of TEM design D4 obtained at 45 °C on a 250 µM solution of protein prepared in MES pH 6.5 with 200 mM NaCl, at 800 MHz ^{1}H frequency.

	D1	D2	D3	D4	TEM-1	SHV-1	LAP-1	LAP-2	KPC-2	CTX-M-1
D1	1	0.875	0.878	0.886	0.551	0.49	0.464	0.46	0.352	0.322
D2	0.875	1	0.905	0.875	0.548	0.487	0.479	0.475	0.36	0.318
D3	0.878	0.905	1	0.867	0.551	0.494	0.475	0.471	0.356	0.341
D4	0.886	0.875	0.867	1	0.536	0.49	0.464	0.46	0.349	0.318
TEM-1	0.551	0.548	0.551	0.536	1	0.655	0.616	0.613	0.402	0.362
SHV-1	0.49	0.487	0.494	0.49	0.655	1	0.606	0.606	0.416	0.374
LAP-1	0.464	0.479	0.475	0.464	0.616	0.606	1	0.996	0.375	0.352
LAP-2	0.46	0.475	0.471	0.46	0.613	0.606	0.996	1	0.371	0.349
KPC-2	0.352	0.36	0.356	0.349	0.402	0.416	0.375	0.371	1	0.474
CTX-M-1	0.322	0.318	0.341	0.318	0.362	0.374	0.352	0.349	0.474	1

Supplementary Table 2 | Sequence identity comparison for TEM-like designs and other class A β-lactamases.



Supplementary Figure 13 | Phylogenetic analysis of TEM-1 like designs within the class A β-lactamases family. Phylogenetic tree of class A β-lactamases available in the Beta-Lactamase Database(reference) and the designed enzymes. The designs cluster into a family (yellow triangle) that is distinct from other β-lactamases of the same class. Interestingly, the TEM family (blue triangle) displays shorter phylogenetic distances to many other individual enzymes and families than to the designs, showing that our method is capable of effectively exploring a new sub-family of proteins.



Supplementary Figure 14 | Analysis of natural against CARBonAra-generated TEM-like sequences. Direct Coupling Analysis (DCA) computed on multiple sequence alignment of natural β-lactamase variants (left, alignment obtained by HHblits as implemented in the Gremlin website using the TEM-1 sequence as seed) and on sequences obtained by recursive imprinting with CARBonAra (right), whereby TEM-1 residues are imprinted consecutively one at a time, sampling amino acids with the confidences from CARBonAra's predictions. In each upper triangle, gray dots indicate residue-residue contacts observed in the X-ray structure of TEM-1 β-lactamase (PDB 1BT5), while in the lower triangle green/red dots denote coevolution pairs that match/don't match contacts in the structures.

Supplementary Dataset 1 | Sequences of TEM-like designs in FASTA format.

>D1

HPEVLKEVKAAEERLGAPVGFIWLDLDTGEVLAAYNPNQYFPMNSTWKVFLVGAVL
HMIDQGKLKLDERVMYSEKDLVPFSPVTSQHLENGMTVAELMWAAVCHVDNTAAN
LLLKLIGGPASLTAFLKDIGDTITRMTHEEPEHNAAVPGSLDDTTTPISMATTLRGLLT
GPILSEESRKFLMDLMRNNQTCGPYFRAALPAGWYMADRCGTGWNGARGIIAALG
PNGKPSVIVVIMTTGSKASIATQAQAIRNIAAAVIKHA

>D2

HPAVLEVVRDAEKRLGAPVGFILLDLETGEVLASYNPNKYFPMCSTWKVFLVGAVL
HMVDQGKLKLDERIMYSEKDLVPFSPVTSQHLENGMTVEELMWAAVCHVDNTAAN
LLLKLIGGPAKLTAFLRDMGDTHTNMTHEEPEHNAAKPGSLDDTSTPISMATTLRGL
LTGPILSEEGRKFLMNLMRNNQVCGPYFRAALPAGWFMADRCGTGWNGARGIVA
ALGPNGKPTQILVIMTTGSKASIEEQHEAIRNIAAAVIKHA

>D3

HPAVLEEVRAAEERLGAPVGFILLDLETGEVLASYNPDKYFPMCSTWKVFLVGAVL
HMVDQGKLKLDERVMYSEEDLVPFSPVTSQHLEDGMTVAELMWAAVCYVDNTAA
NLLLKLIGGPAKLTAFLRDMGDTVTNMTHMEPEHNAAVPGSLDDTTTPISMATTLR
GLLTGPILSEEGRKFLSDLMRNNQHCGPYFRAALPAGWYMADRCGSGWNGARGI
VAAFGPNGKPSQIVVYMTTGSKASIEERHQCIRNIAAAVIKHA

>D4

HPEVLKVVRAAEERLGAPVGFIWLDLDTGEVLAAYNPDKYFPMCSTWKVFLVGAVL
HMIDQGKLKRDERIMYSEKDLVPFSPVCSQHLENGMTVEELMWAAVCYVDNTAAN
LLLKLIGGPAKLTAFLRDIGDTVNRMTHEEPDHNAAEPGSLDDTTTPISMATTLRGLL
TGPILSEEARKFLQDLMANNQYCGPYFRAALPAGWFLADRCGSGWNGARGIVAAL
GPNGKPSVIVVIMTSGSTASMETQHEAIRNIAAAVIKHA