LEARNING FROM PROTEIN STRUCTURE WITH GEOMETRIC VECTOR PERCEPTRONS

Bowen Jing , Stephan Eismann , Patricia Suriana, Raphael J.L. Townshend, Ron O. Dror

Stanford University

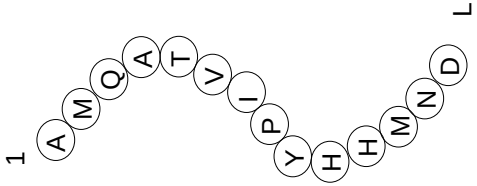
Presented by: Md Hossain Shuvo

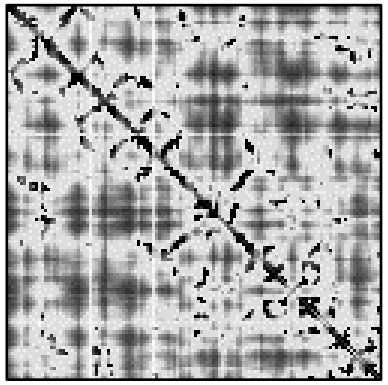
Virginia Tech

Motivation

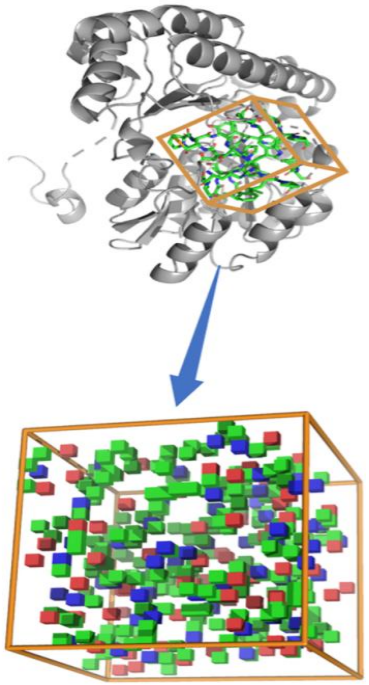
Learning from Protein Structure

Sequential representation

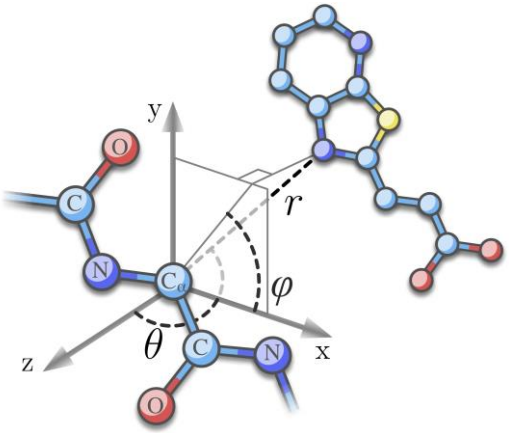




Voxelized representation



Graph structure representation

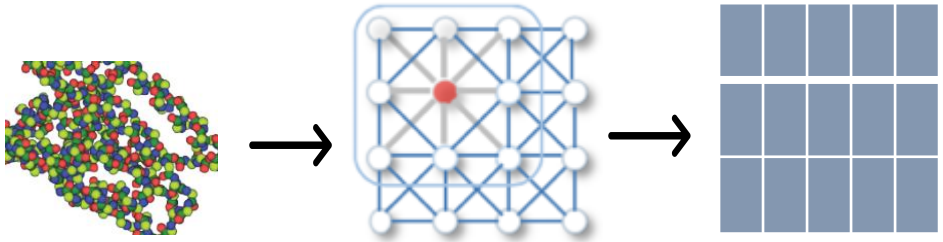


src: https://team.inria.fr/nano-d/software/korp-pl/
https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0221347.g002

Motivation

Learning from Protein Structure

Learning with CNN



Geometric aspect

Learning with GNN

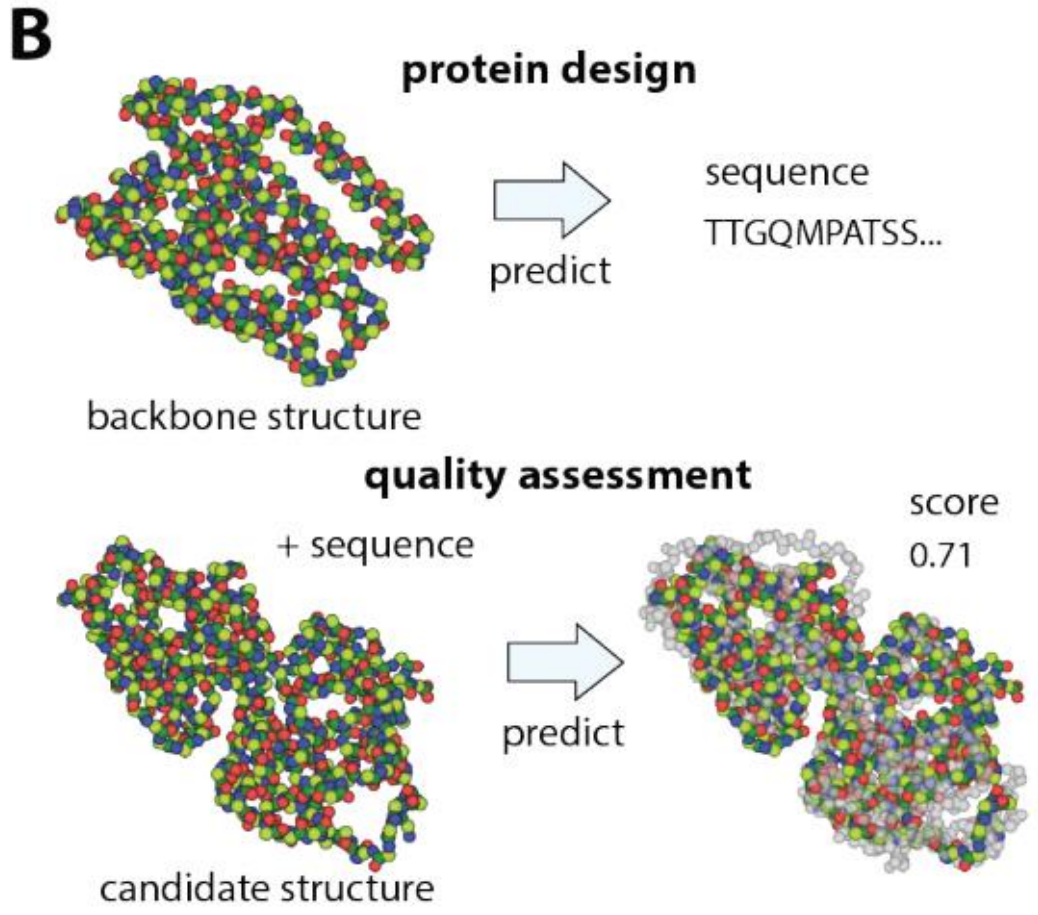
Relational aspect

Geometric Vector Perceptrons (GVP)

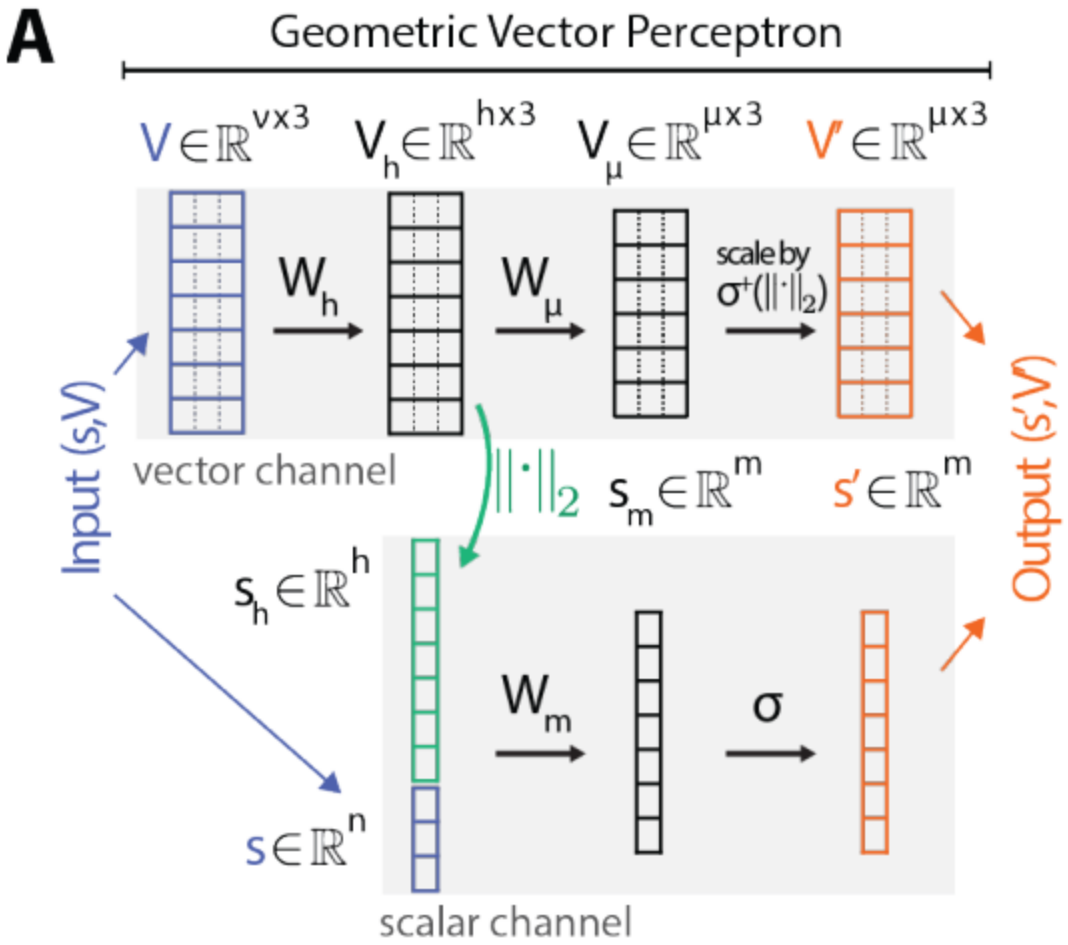
Background

Geometric Vector Perceptrons (GVP)

- GPV layer
- Improves GNN
- MLP → GVP in GNN
- Operates directly on scalar and geometric features



Architecture



Input: Scalar and vector features $(\mathbf{s}, \mathbf{V}) \in \mathbb{R}^n \times \mathbb{R}^{\nu \times 3}$.
Output: Scalar and vector features $(\mathbf{s}', \mathbf{V}') \in \mathbb{R}^m \times \mathbb{R}^{\mu \times 3}$.

$h \leftarrow \max(\nu, \mu)$

GVP:

$\mathbf{V}_h \leftarrow \mathbf{W}_h \mathbf{V} \quad \in \mathbb{R}^{h \times 3}$

$\mathbf{V}_\mu \leftarrow \mathbf{W}_\mu \mathbf{V}_h \quad \in \mathbb{R}^{\mu \times 3}$

$\mathbf{s}_h \leftarrow \|\mathbf{V}_h\|_2 \text{ (row-wise)} \quad \in \mathbb{R}^h$

$\mathbf{v}_\mu \leftarrow \|\mathbf{V}_\mu\|_2 \text{ (row-wise)} \quad \in \mathbb{R}^\mu$

$\mathbf{s}_{h+n} \leftarrow \text{concat}(\mathbf{s}_h, \mathbf{s}) \quad \in \mathbb{R}^{h+n}$

$\mathbf{s}_m \leftarrow \mathbf{W}_m \mathbf{s}_{h+n} + \mathbf{b} \quad \in \mathbb{R}^m$

$\mathbf{s}' \leftarrow \sigma(\mathbf{s}_m) \quad \in \mathbb{R}^m$

$\mathbf{V}' \leftarrow \sigma^+(\mathbf{v}_\mu) \odot \mathbf{V}_\mu, \text{ (row-wise multiplication)} \quad \in \mathbb{R}^{\mu \times 3}$

return $(\mathbf{s}', \mathbf{V}')$

Dataset

Protein design

- CATH4.2 dataset
- 18204, 608 and 1120 structures for training, validation and testing, respectively
- TS50 dataset for testing
- Sequence identity < 30%

MQA

Training:

CASP5- 10: 79200 models for 528 targets

Testing:

CASP11 -12: 84 and 40 targets respectively (stage1 & 2)

CASP13: 20 targets (stage2)

Features

Node features

- Scalar features $\{\sin, \cos\} \circ \{\phi, \psi, \omega\}$, where ϕ, ψ, ω are the dihedral angles computed from C_{i-1}, N_i, $C\alpha_i$, C_i, and N_{i+1}.
- The *forward* and *reverse* unit vectors in the directions of $C\alpha_{i+1} - C\alpha_i$ and $C\alpha_{i-1} - C\alpha_i$, respectively.
- The unit vector in the imputed direction of $C\beta_i - C\alpha_i$.[3] This is computed by assuming tetrahedral geometry and normalizing

$$\sqrt{\frac{1}{3}}(\mathbf{n} \times \mathbf{c})/||\mathbf{n} \times \mathbf{c}||_2 - \sqrt{\frac{2}{3}}(\mathbf{n} + \mathbf{c})/||\mathbf{n} + \mathbf{c}||_2$$

  where $\mathbf{n} = N_i - C\alpha_i$ and $\mathbf{c} = C_i - C\alpha_i$. This vector, along with the forward and reverse unit vectors, unambiguously define the orientation of each amino acid residue.
- A one-hot representation of amino acid identity, when available.

Features

Edge features

The set of edges is $\mathcal{E} = \{\mathbf{e}_{j\to i}\}_{i\neq j}$ for all i, j where $\mathfrak{v}_j$ is among the $k = 30$ nearest neighbors of $\mathfrak{v}_i$ as measured by the distance between their Cα atoms. Each edge has an embedding $\mathbf{h}_e^{(j\to i)}$ with the following features:

- The unit vector in the direction of $\text{C}\alpha_j - \text{C}\alpha_i$.
- The encoding of the distance $||\text{C}\alpha_j - \text{C}\alpha_i||_2$ in terms of Gaussian radial basis functions.[4]
- A sinusoidal encoding of $j - i$ as described in Vaswani et al. (2017), representing distance along the backbone.

Model training

Protein design

- Network learns a generative model
- Models the distribution for each specific position
- Outputs the 20-way probability

MQA

- Regression against the true quality score (GDT-TS)

Learning rate	10^{-4} to 10^{-3}
Dropout probability	10^{-4} to 10^{-1}
Number of graph layer	3 to 6
MQA pairwise loss	0 – 2
Epochs	100
Optimizer	Adam
Input dimension	**Node:** 16 vectors 100 channels **Edge:** 1 vector 32 channel
Batch size	1800 and 3000 residues for CPD and MQA, respectively

Performance evaluation

Computational Protein design (CPD)

- Perplexity
  - Lower is better
- Recovery: mean recovery of 100 sequences
  - Higher is better

Model quality estimation (MQA)

- Global and per-target Pearson correlation coefficients
  - GDT-TS score
  - Higher is better

Performance evaluation: CPD

		Perplexity			Recovery %		
Method	Type	Short	Single-chain	All	Short	Single-chain	All
GVP-GNN	GNN	**7.10**	**7.44**	**5.29**	**32.1**	**32.0**	**40.2**
Structured GNN	GNN	8.31	8.88	6.55	28.4	28.1	37.3
Structured Transformer	GNN	8.54	9.03	6.85	28.3	27.6	36.4

Table 3: Performance on the CATH 4.2 test set and its short and single-chain subsets in terms of per-residue perplexity (lower is better) and recovery (higher is better). Recovery is reported as the median over all structures of the mean recovery of 100 sequences per structure. GVP-GNN performs better than Structured Transformer and a variant of it, Structured GNN, in which we replaced the attention mechanisms with standard graph propagation operations (see main text).

Performance evaluation: MQA

		CASP 11				CASP 12			
		Stage 1		Stage 2		Stage 1		Stage 2	
Method	Type	Glob	Per	Glob	Per	Glob	Per	Glob	Per
GVP-GNN	GNN	**0.84**	**0.66**	**0.87**	**0.45**	**0.79**	**0.73**	**0.82**	**0.62**
3DCNN	CNN	0.59	0.52	0.64	0.40	0.49	0.44	0.61	0.51
Ornate	CNN	0.64	0.47	0.63	0.39	0.55	0.57	0.67	0.49
GraphQA	GNN	0.83	0.63	0.82	0.38	0.72	0.68	0.81	0.61
VoroMQA	Seq	0.69	0.62	0.65	0.42	0.46	0.61	0.61	0.56
SBROD	Seq	0.58	0.65	0.55	0.43	0.37	0.64	0.47	0.61
ProQ3D	Seq	0.80	**0.69**	0.77	0.44	0.67	0.71	0.81	0.60

CASP13

Method	Global	Per-target
GVP-GNN	**0.888**	**0.671**
SASHAN	0.840	0.633
FaeNNz	0.810	0.650
VoroMQA-A	0.744	0.595
VoroMQA-B	0.726	0.586
ProQ3D	0.847	0.660
MULTICOM-NOVEL	0.652	0.551
ProQ4	0.604	**0.691**

Ablation study

	MQA				CPD	
	CASP 11 Stage 2		CASP 12 Stage 2		CATH4.2 All	
Modification	Global	Per-target	Global	Per-target	Perplexity	Recovery
None	**0.87**	**0.45**	0.82	**0.62**	**5.29**	**40.2**
MLP layer	0.84	0.36	0.79	0.59	7.76	30.6
Only scalars	0.84	0.38	**0.83**	0.59	7.31	32.4
Only vectors	0.56	0.16	0.57	0.39	11.05	23.2
No $\mathbf{W}_\mu$	0.86	0.41	0.81	0.60	5.85	37.1
GraphQA	0.82	0.38	0.81	0.61	–	–
Structured GNN	–	–	–	–	6.55	37.3

Conclusion

- GVP-GNN to learn both relational and geometric representations
- Enhance the expressive power of GNN
- Posses the equivariant and invariance properties
- Demonstrated on MQA and CPD problems
- Future application: protein complexes, RNA structure and protein-ligand interactions
- Code published and available at
  - https://github.com/drorlab/gvp