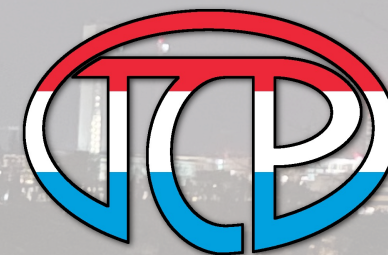




UNIVERSITÉ DU
LUXEMBOURG

University of Luxembourg
Department of Physics and Materials Science
Theoretical Chemical Physics Group
Advisor: Alexandre Tkatchenko
Adil Kabylda / kabylda.github.io



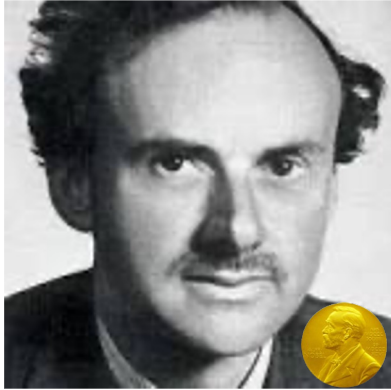
Quantum (Bio)Molecular Simulations with Machine Learning Force Fields

PASC26, From Physics to Data - and Back, 29th June 2026

Outline

- ***Introduction***
- *SO3LR: General-purpose MLFF for organic matter*
- *QCell: QM dataset spanning diverse biomolecular fragments*
- *Conclusions & Outlook*

From First Principles to Complex Matter



P. A. M. Dirac

Quantum mechanics of many-electron Systems
Proc. R. Soc. Lond., 123, 714 (1929)



E. Schrödinger

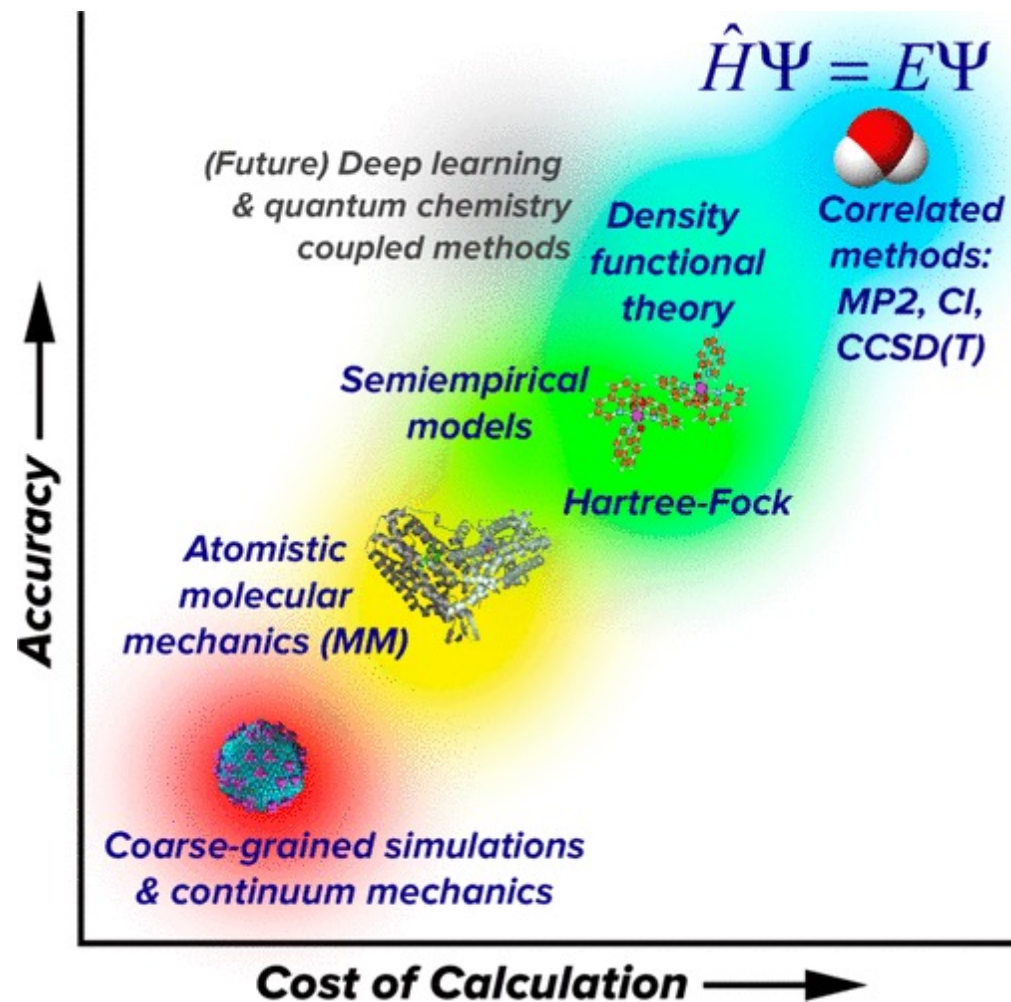
What is Life? The Physical Aspect of the Living Cell
(Cambridge University Press, 1944)



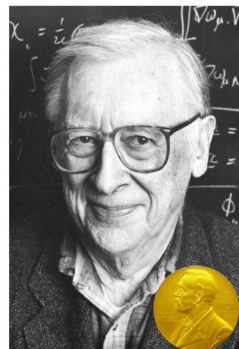
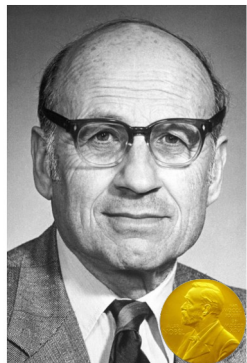
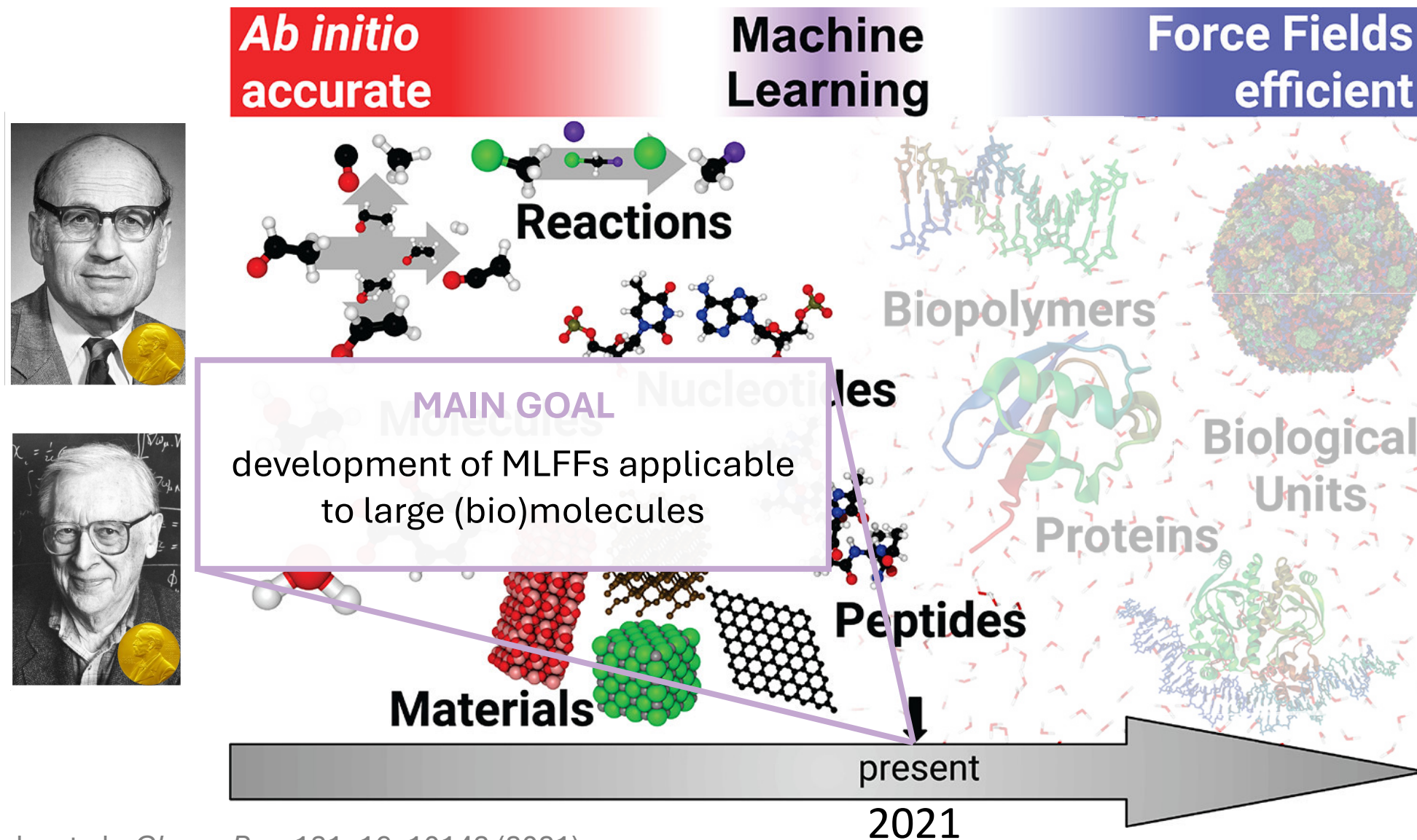
R. P. Feynman, R. B. Leighton, and M. Sands
The Feynman lectures on physics, Volume I.

- Realizing Schrödinger's dream requires development of approximate practical methods

Hierarchy of Modelling Methods

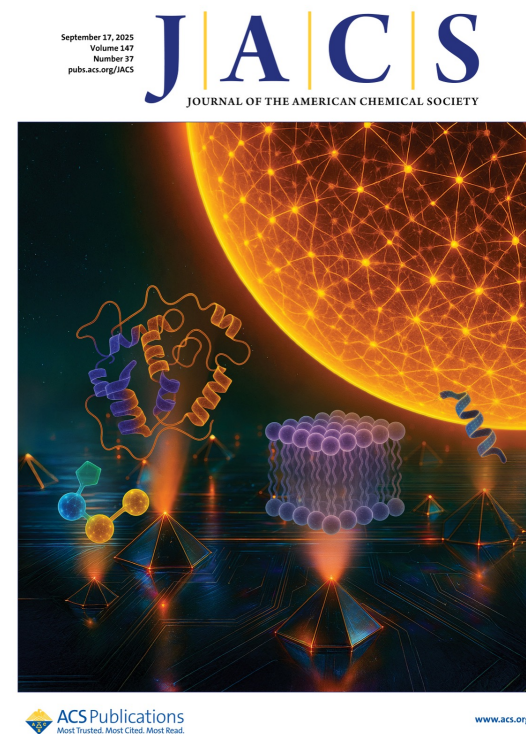


Machine Learning Force Fields (MLFFs)



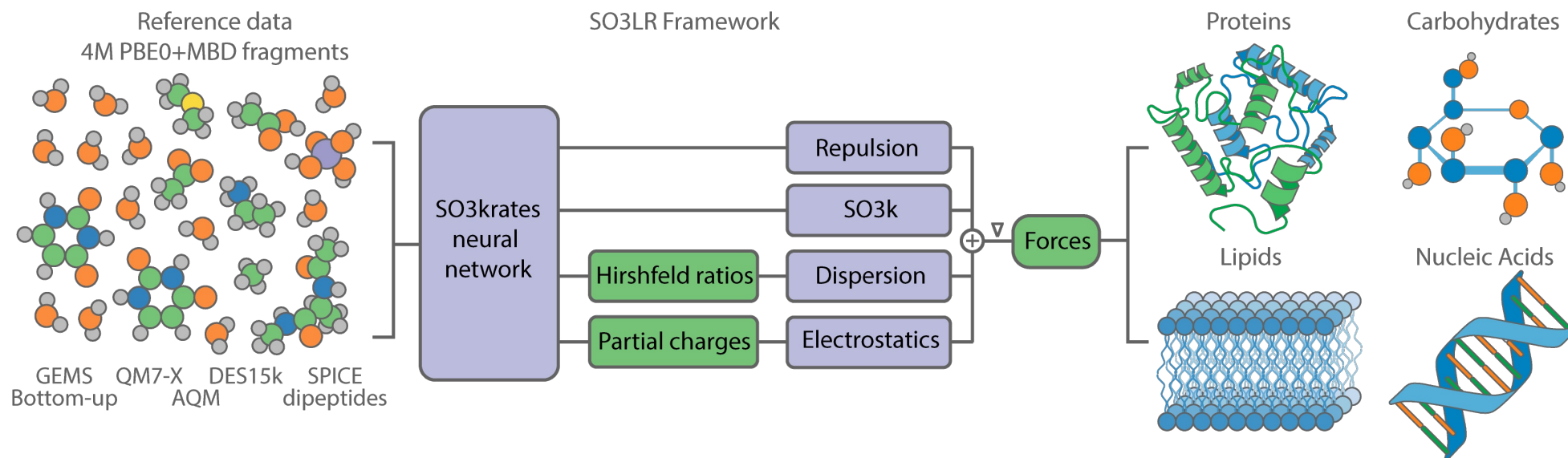
Outline

- *Introduction*
- **SO3LR: General-purpose MLFF for organic matter**
- *QCell: QM dataset spanning diverse biomolecular fragments*
- *Conclusions & Outlook*



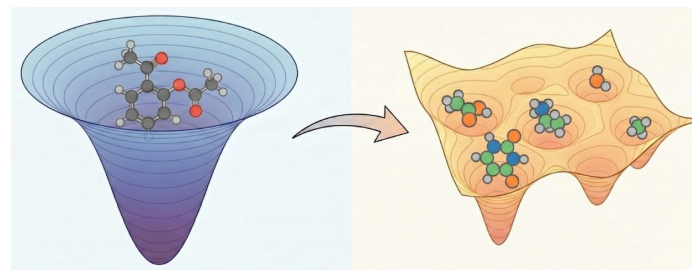
A. Kabylda, J. T. Frank, S. Suárez-Dou, A. Khabibrakhmanov, L. Medrano Sandonas, O. T. Unke, S. Chmiela, K.-R. Müller, A. Tkatchenko
Molecular Simulations with a Pretrained Neural Network and Universal Pairwise Force Fields
Journal of the American Chemical Society 147, 37, 33723 (2025). Cover article.

SO3LR: General-purpose MLFF for organic matter

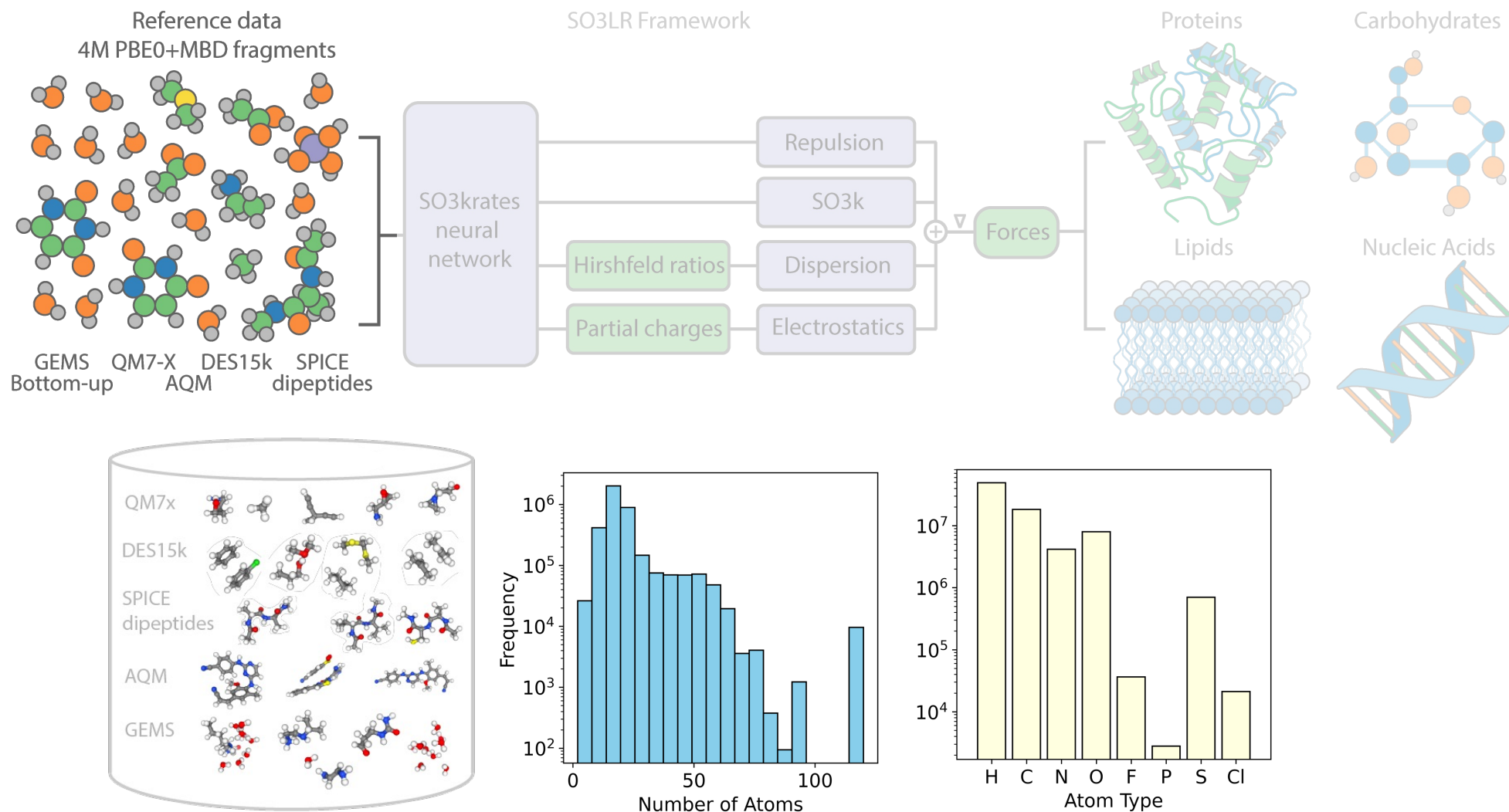


- MLFF architectures
- Quantum-mechanical datasets
- Physical interaction models
- GPU software/hardware

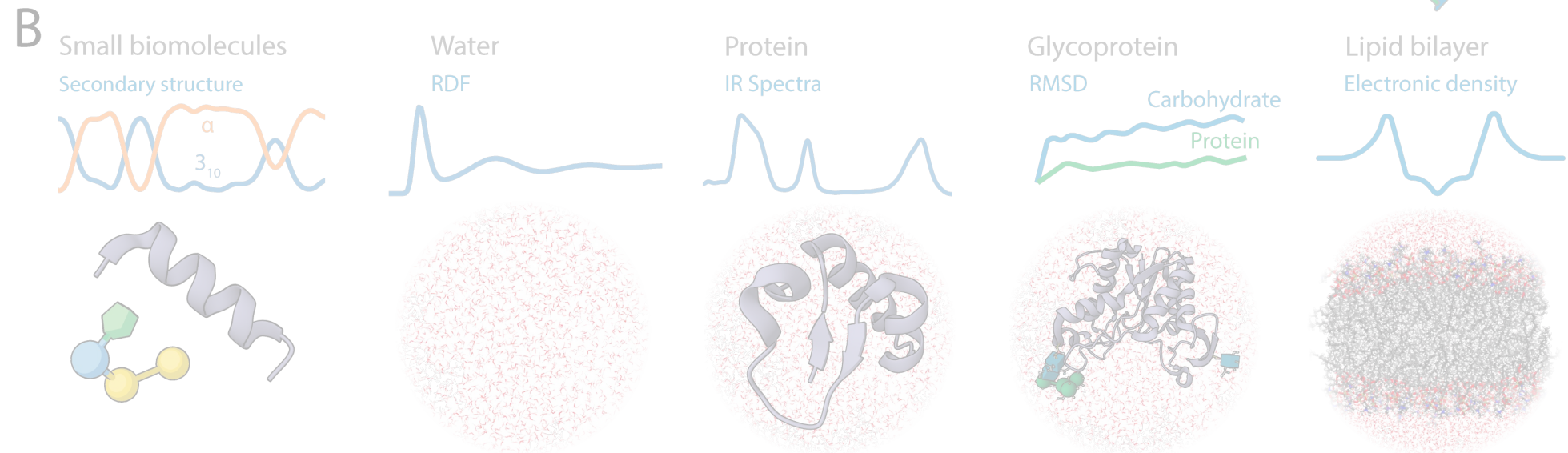
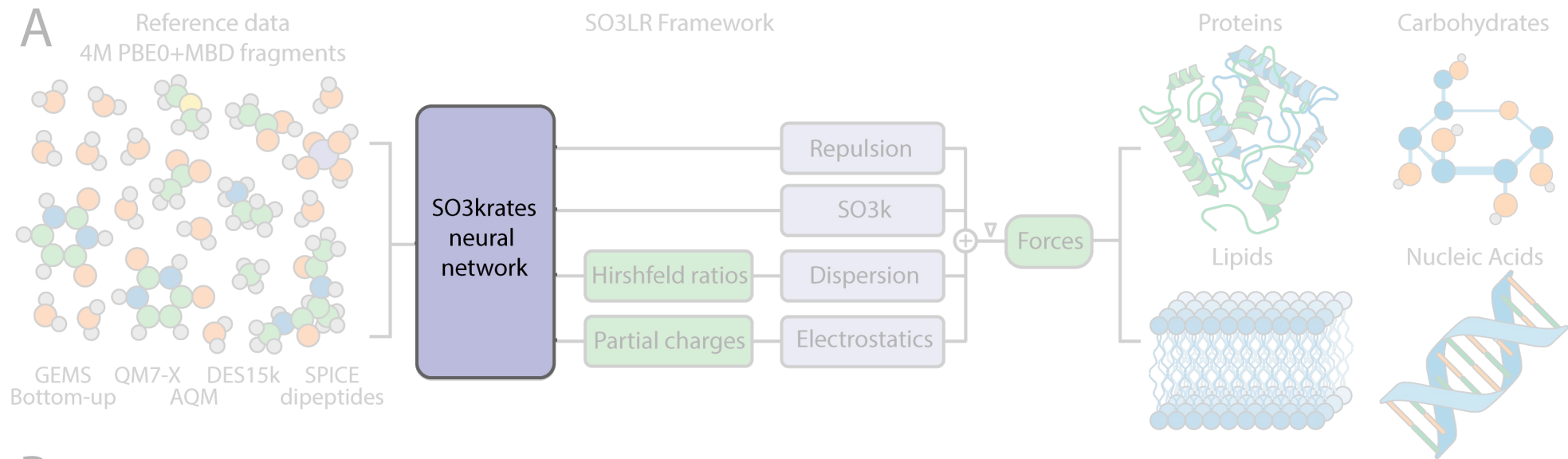
*System-specific -> General-Purpose
MLFFs*



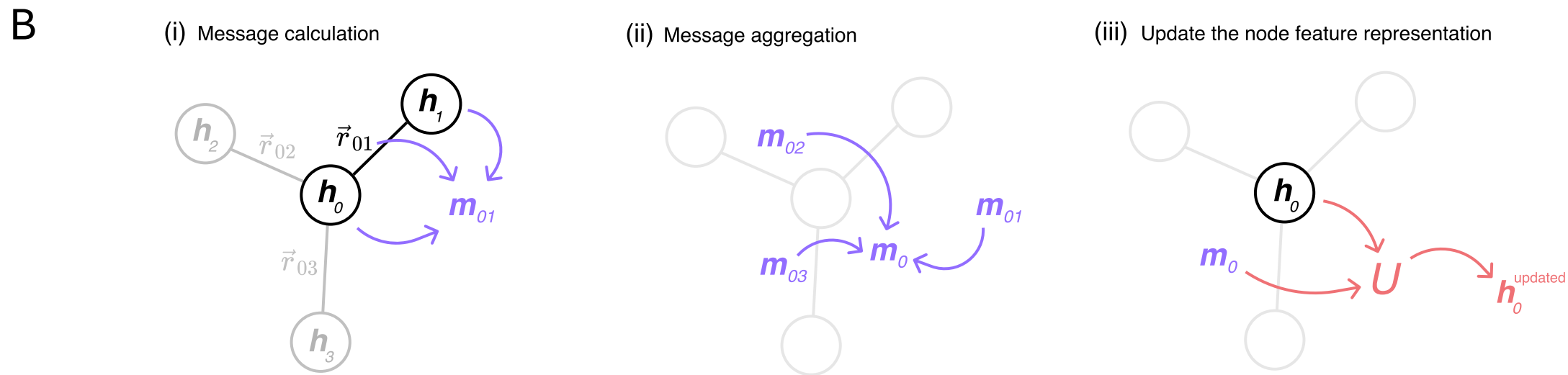
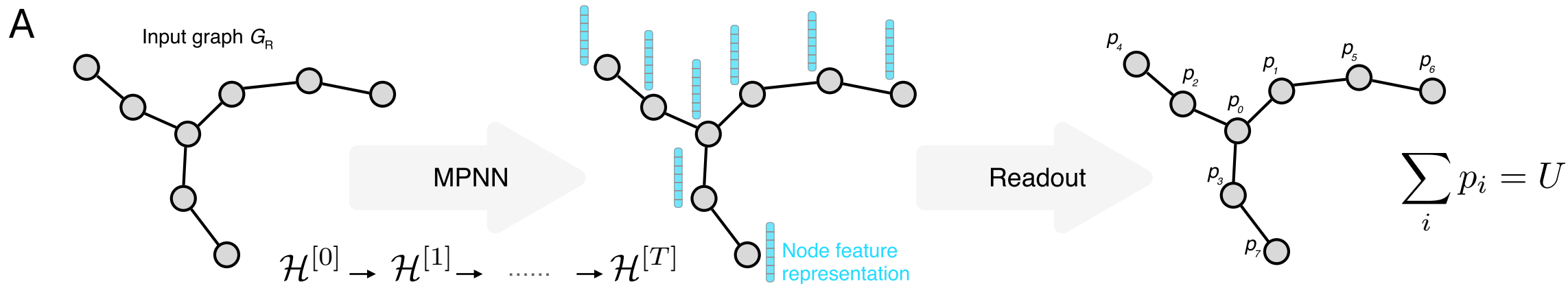
Diverse QM training dataset



SO3LR



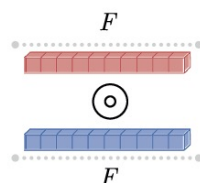
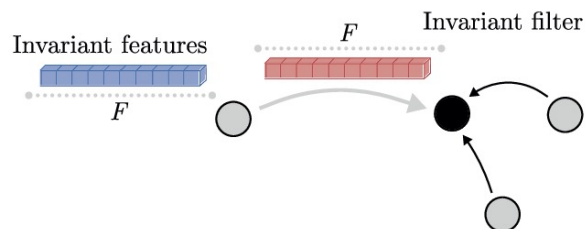
Message Passing Neural Networks



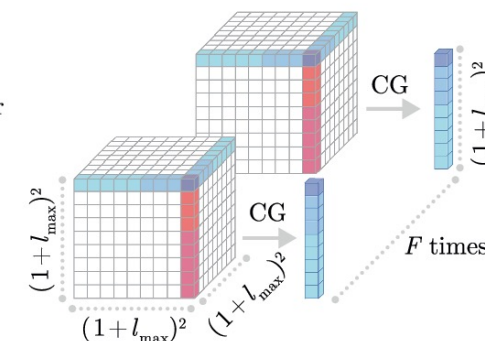
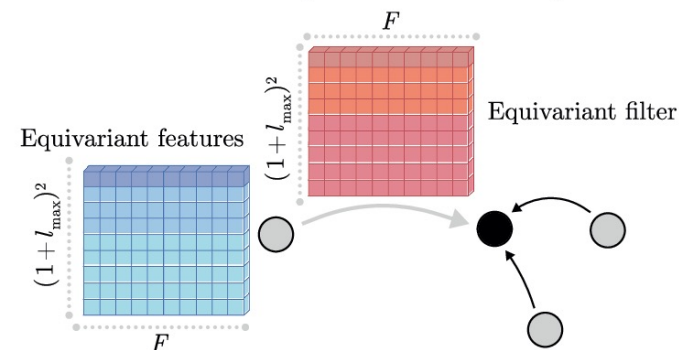
SO3krates – cornerstone of SO3LR

(a) **Invariant convolution** $\left[\mathcal{O}(F)\right]$

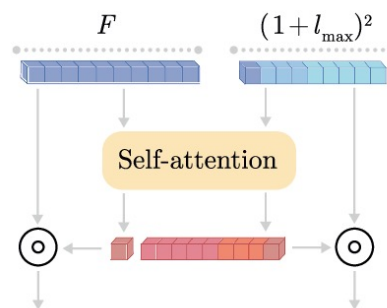
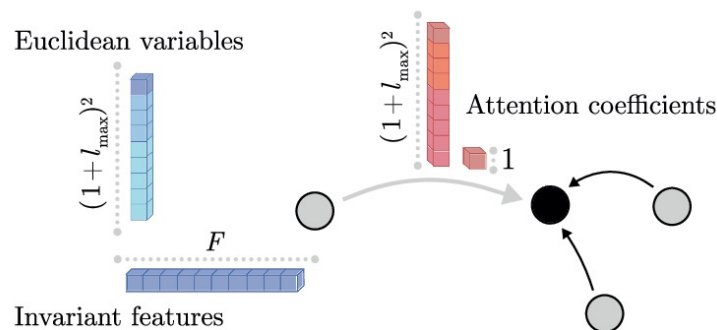
F : Feature dimension $l_{\max}$: Maximal degree
CG: Clebsch-Gordan Contraction



(b) **SO(3) convolution** $\left[\mathcal{O}(F \times (l_{\max} + 1)^6)\right]$

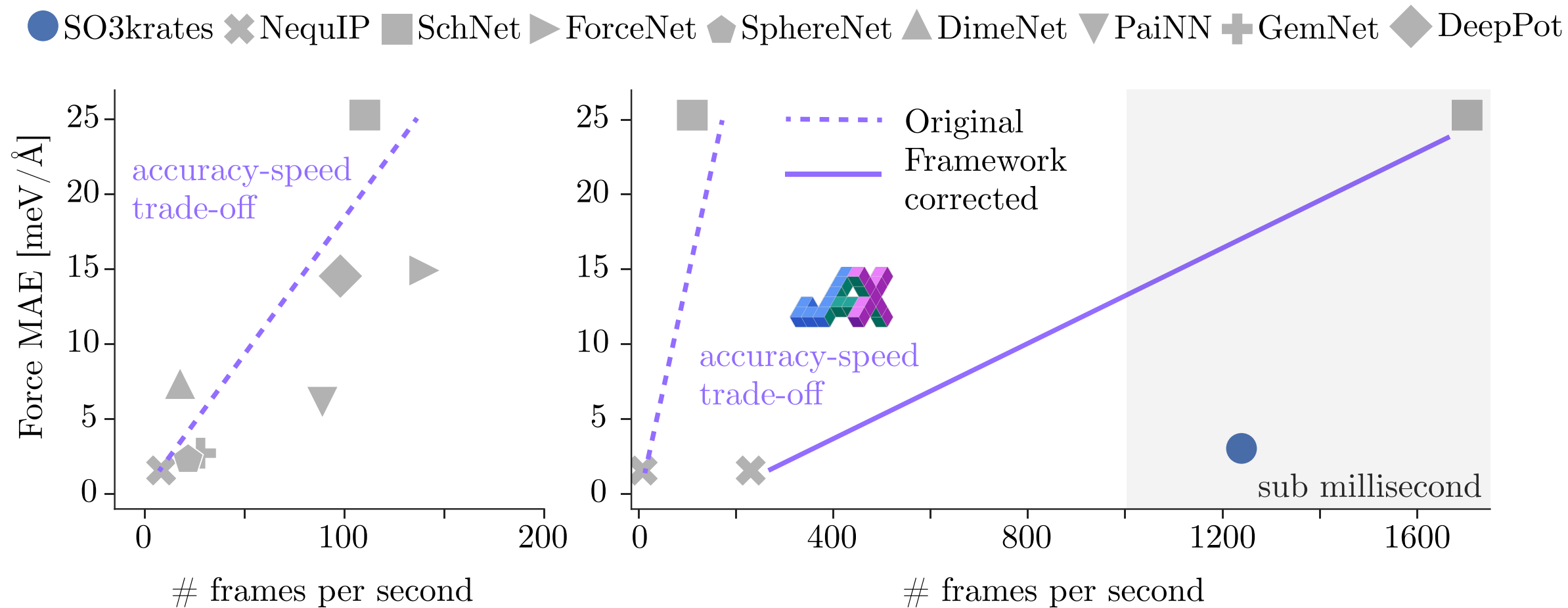


(c) **Euclidean Attention** $\left[\mathcal{O}(F + (l_{\max} + 1)^2)\right]$

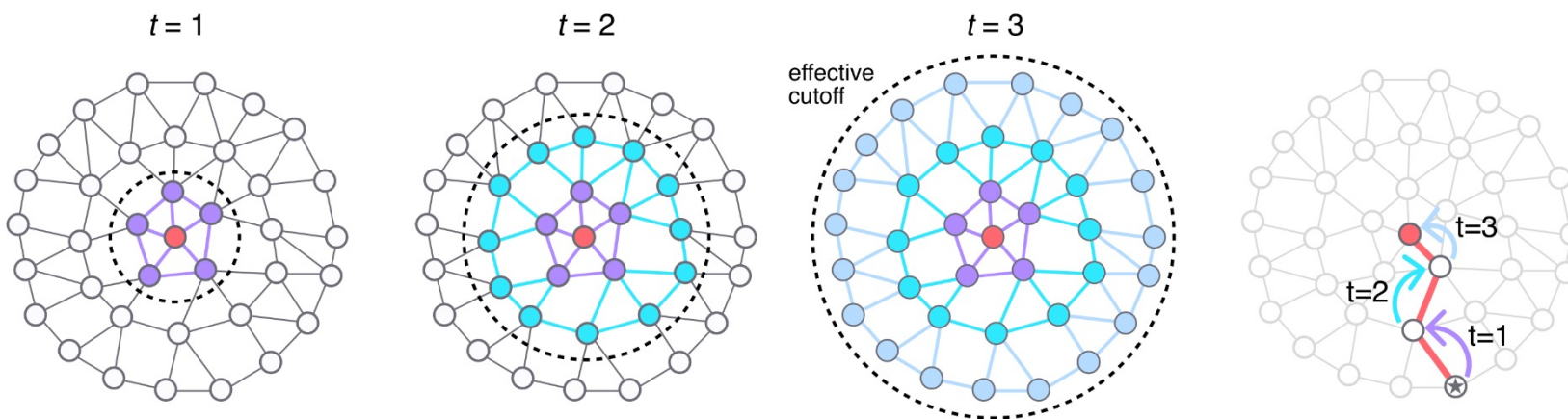
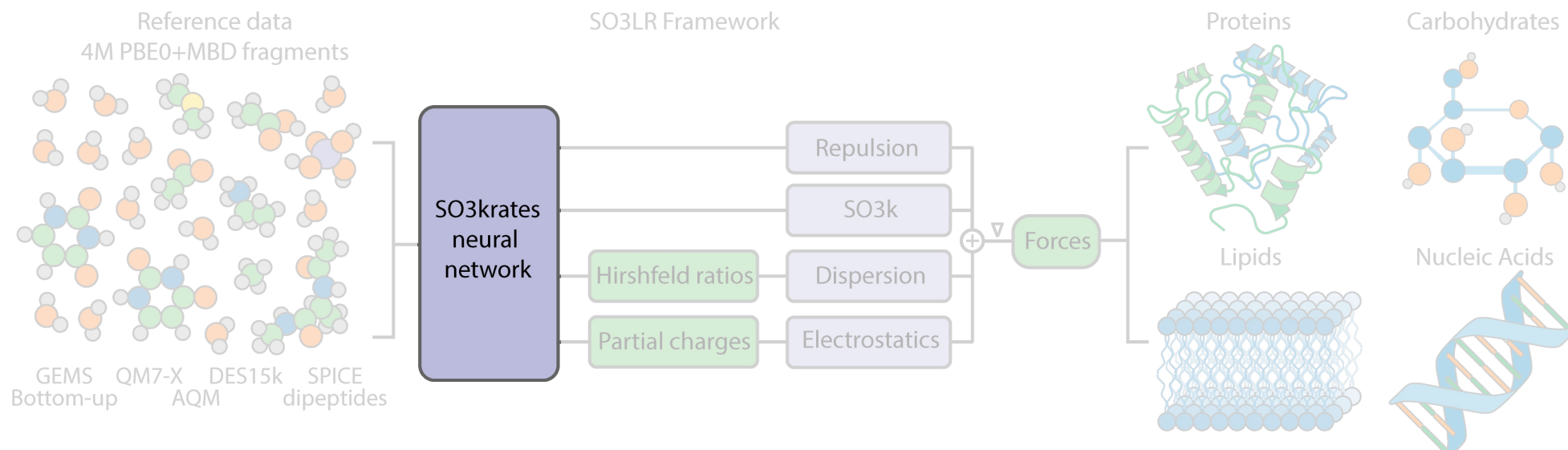


- Equivariant MPNN with two conceptual changes:
1. separates the message into an invariant and an equivariant part
 2. replaces the SO(3) convolution by an attention function on its invariant output

MLFFs accuracy-speed trade-off



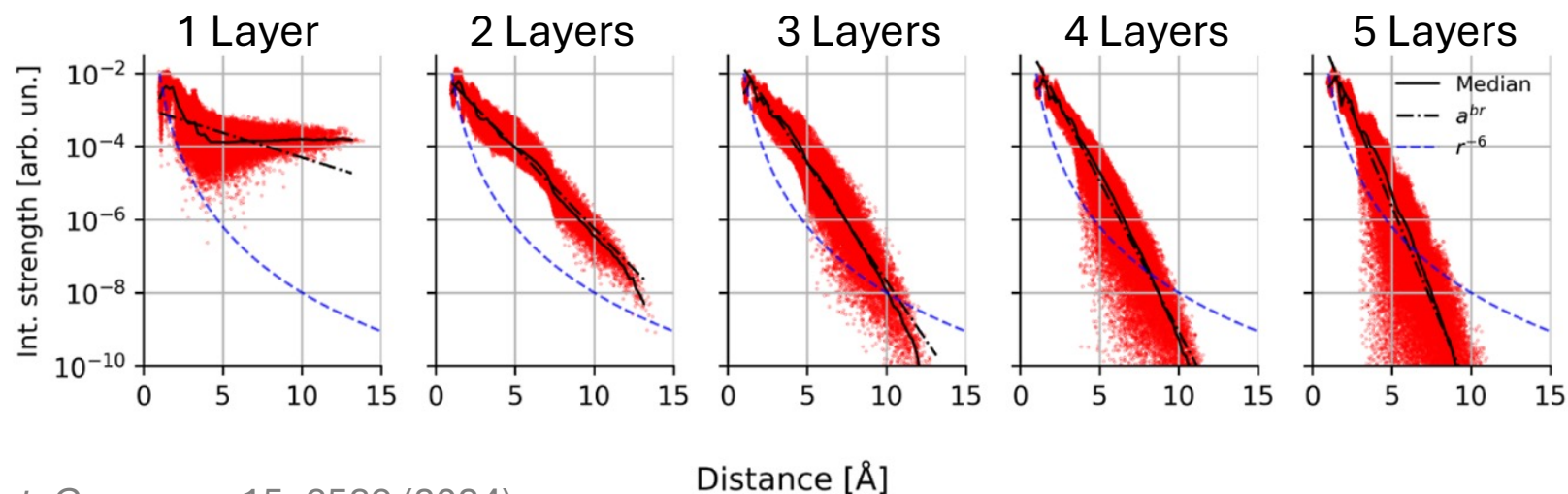
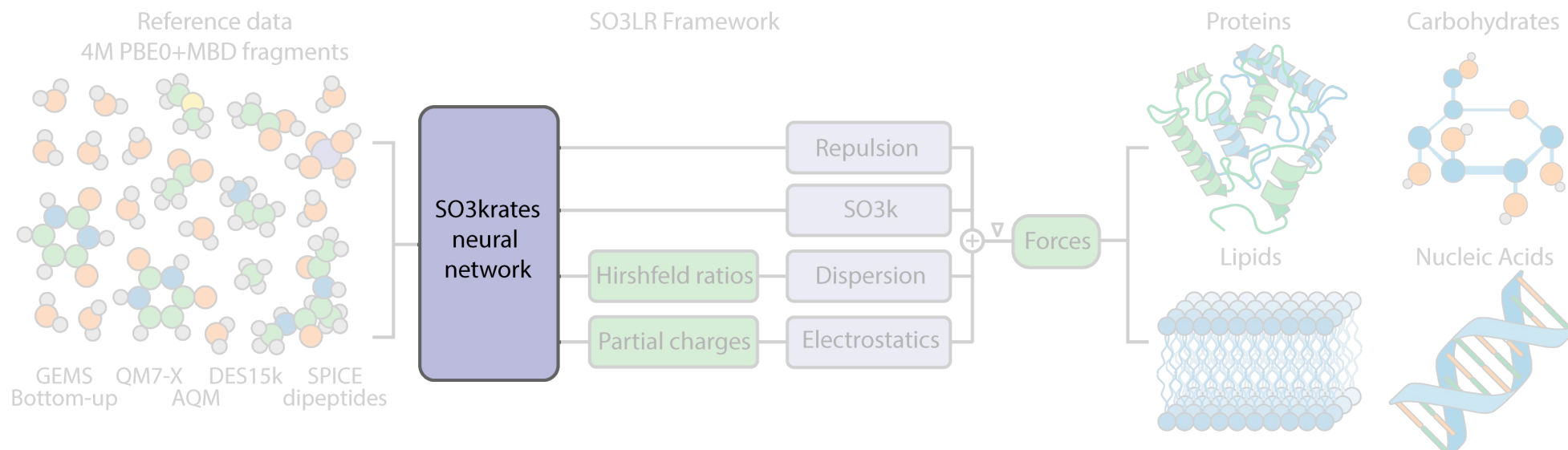
SO3krates – cornerstone of SO3LR



Frank et al., *Nat. Commun.* 15, 6539 (2024)

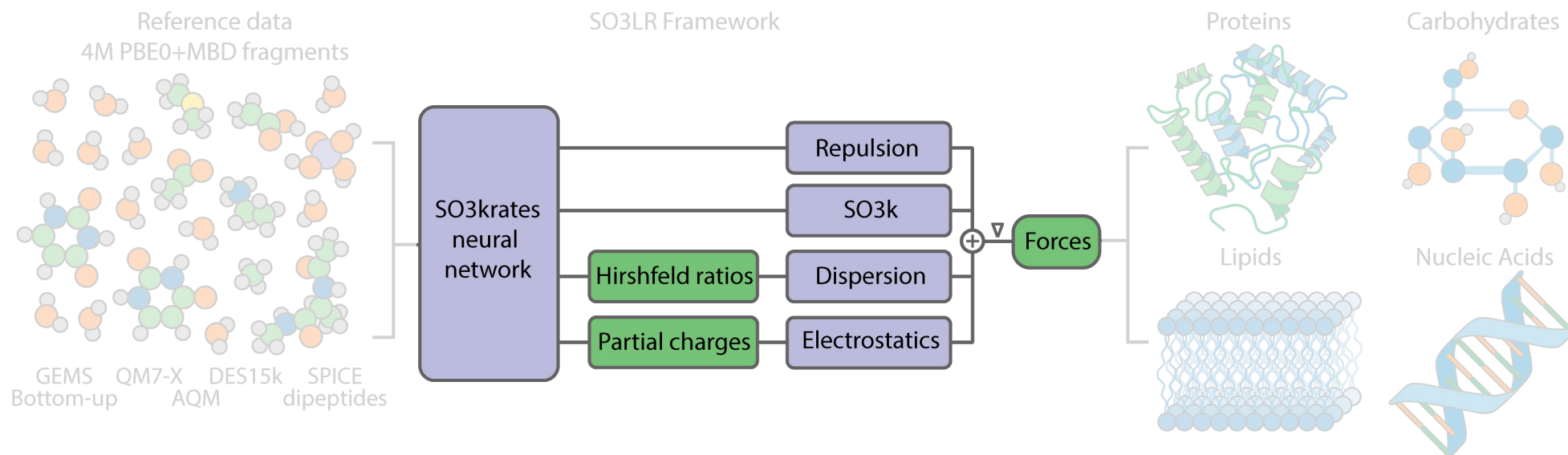
Esders, Schnake, Lederer, Kabylda et al., *JCTC* 21, 714 (2025)

SO3krates – cornerstone of SO3LR



- MPNNs struggle to describe long-range interactions

Universal Pairwise Force Fields



$$E_{\text{Pot}} = \underbrace{E_{\text{ZBL}}}_{\text{short-range}} + \underbrace{E_{\text{SO3k}}}_{\text{semi-local}} + \underbrace{E_{\text{Elec}} + E_{\text{Disp}}}_{\text{long-range}}$$

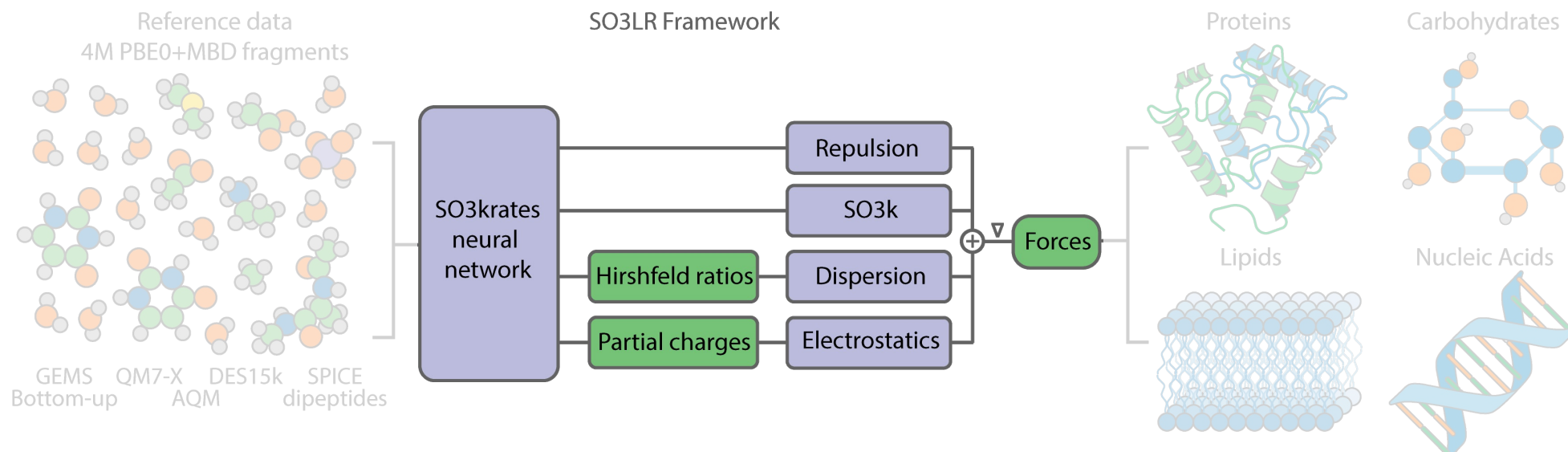
$$E_{\text{Elec}} = \sum_{i < j} q_i q_j \frac{\text{erf}(r_{ij}/\sigma)}{r_{ij}}$$

$$E_i, q_i, h_i = \text{SO3krates}(\mathcal{R}, \mathcal{Z}, Q_{\text{tot}}, S_{\text{tot}})$$

$$E_{\text{Disp}} = - \sum_{i < j} \sum_{n=3}^5 \frac{C_{2n}^{ij}}{r_{ij}^{2n} + R_{d,ij}^{2n}}$$

- Charges and Hirshfeld ratios ($V_{\text{eff}}/V_{\text{free}}$) are environment-dependent

SO3LR Training



Loss function:

$$\mathcal{L} = \frac{\lambda_F}{B} \sum_{b=1}^B \frac{1}{N_b} \sum_{i=1}^{N_b} \|\vec{F}_{i,\text{true}} - \vec{F}_{i,\text{pred}}\|_2^2$$

$$+ \frac{\lambda_\mu}{B} \sum_{b=1}^B \|\vec{\mu}_{b,\text{true}} - \vec{\mu}_{b,\text{pred}}\|_2^2$$

$$+ \frac{\lambda_h}{B} \sum_{b=1}^B \frac{1}{N_b} \sum_{i=1}^{N_b} (h_{i,\text{true}} - h_{i,\text{pred}})^2,$$

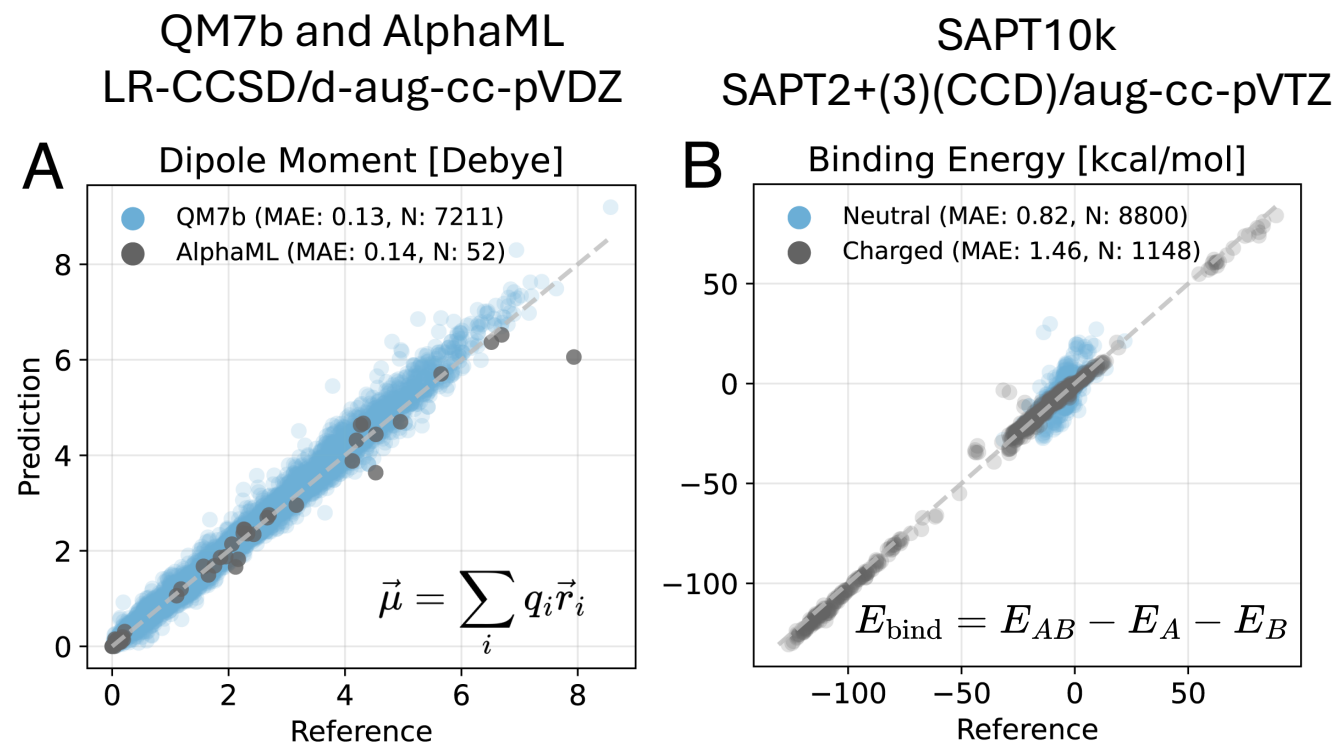
Static benchmarks

TABLE I. Root Mean Square Error of the model on various test sets. Force (eV/Å), dipole moment vector ($e \times \text{\AA}$), and Hirshfeld ratios. Dash indicates no data.

Dataset	Size	# atoms	Force	Dipole	Hirsh. rat.
QM7-X	10000	6-23	0.069	0.031	0.012
GEMS bottom-up	10000	2-120	0.086	0.048	-
AcAla ₃ NMe	100	42	0.052	0.051	0.012
DHA	100	56	0.053	0.072	0.012
AT-AT	100	60	0.168	0.238	0.025
Stachyose	100	87	0.105	0.119	0.016
Buckyball Catcher	100	148	0.384	4.030	0.029
Nanotube	100	370	0.717	2.950	0.039
AcAla ₁₅ NMe	312	162	0.055	-	-
Crambin top-down	5624	230-321	0.057	-	-
TorsionNet500	12000	13-37	0.088	0.061	0.019

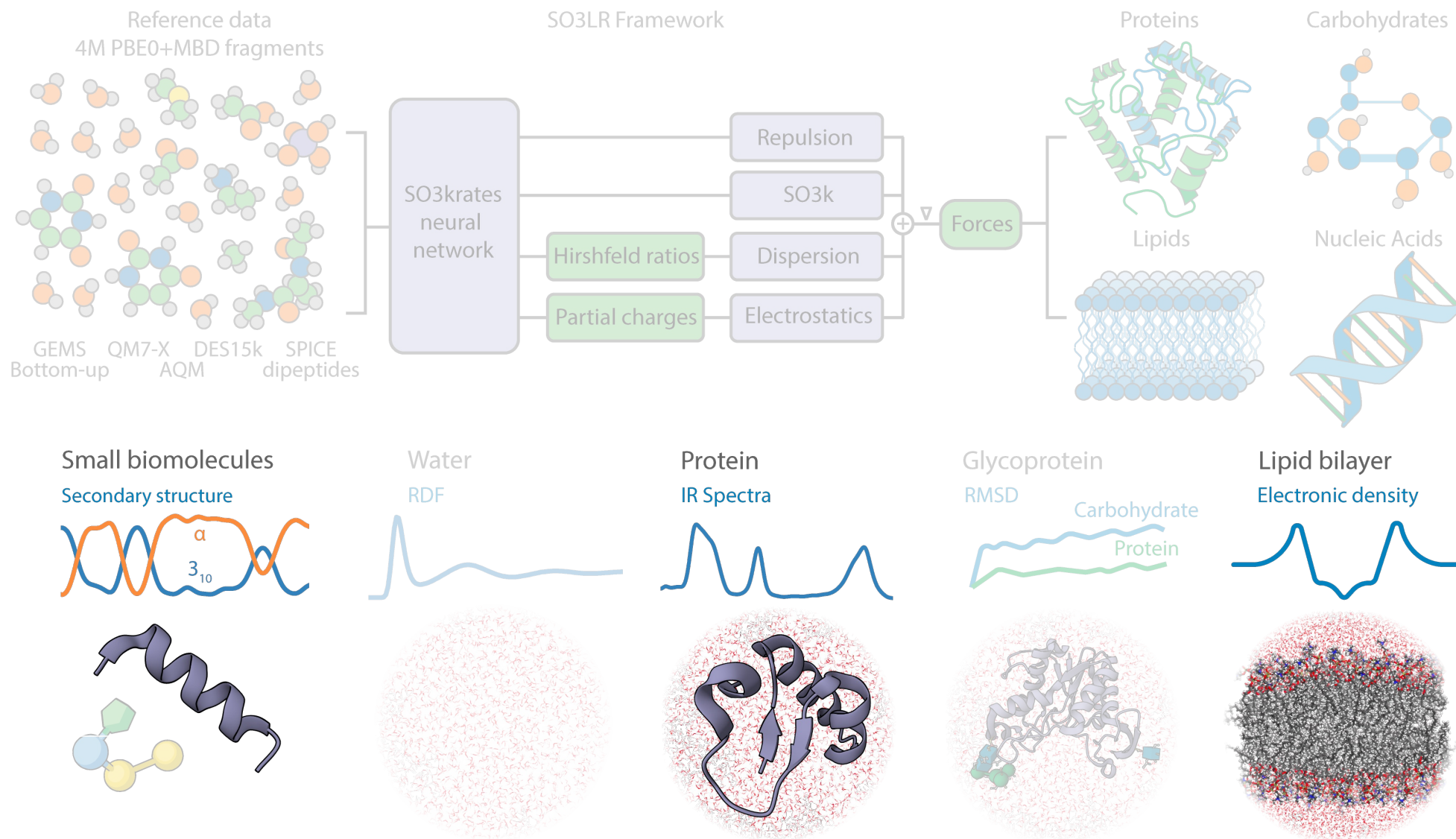
Yang et al., *Sci. Data* 6, 152 (2019)

Villot and Lao, *J. Chem. Phys.* 160, 184103 (2024)

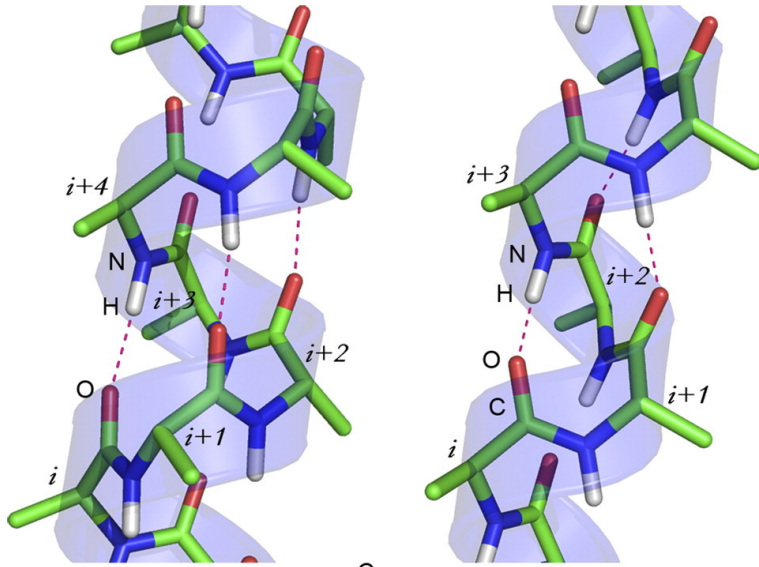


- Accurate dipole moment and binding energy predictions
- Rare outliers include dimers like ClF, P(CNO)₃, and PH₂NO₂ complexes

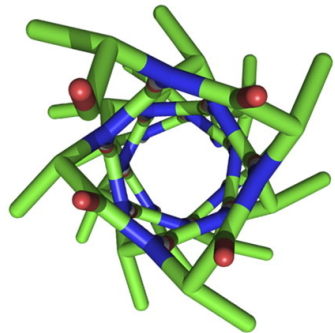
(Bio)Molecular Simulations with SO3LR



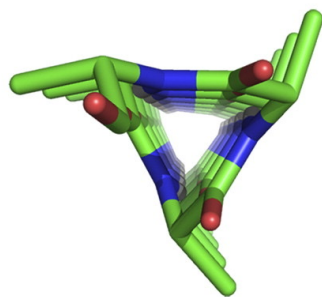
Folding of polyalanine



c

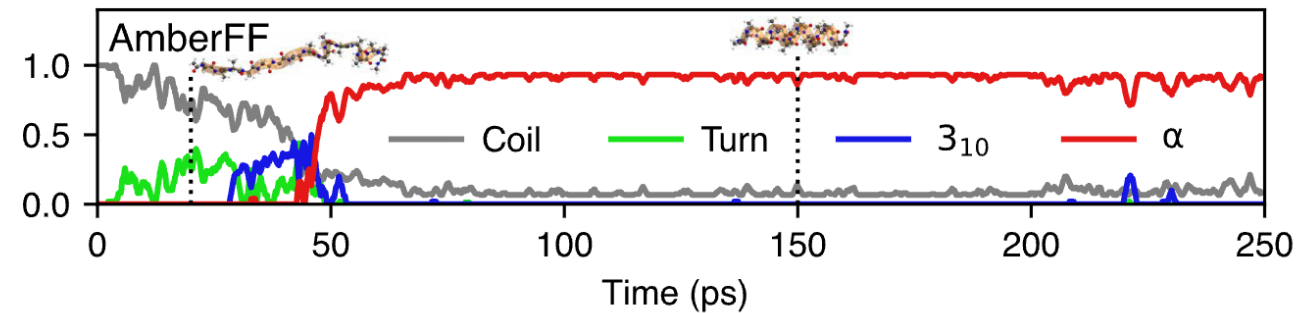


Canonical α -helix



Canonical 3(10) helix

AcAla₁₅NMe

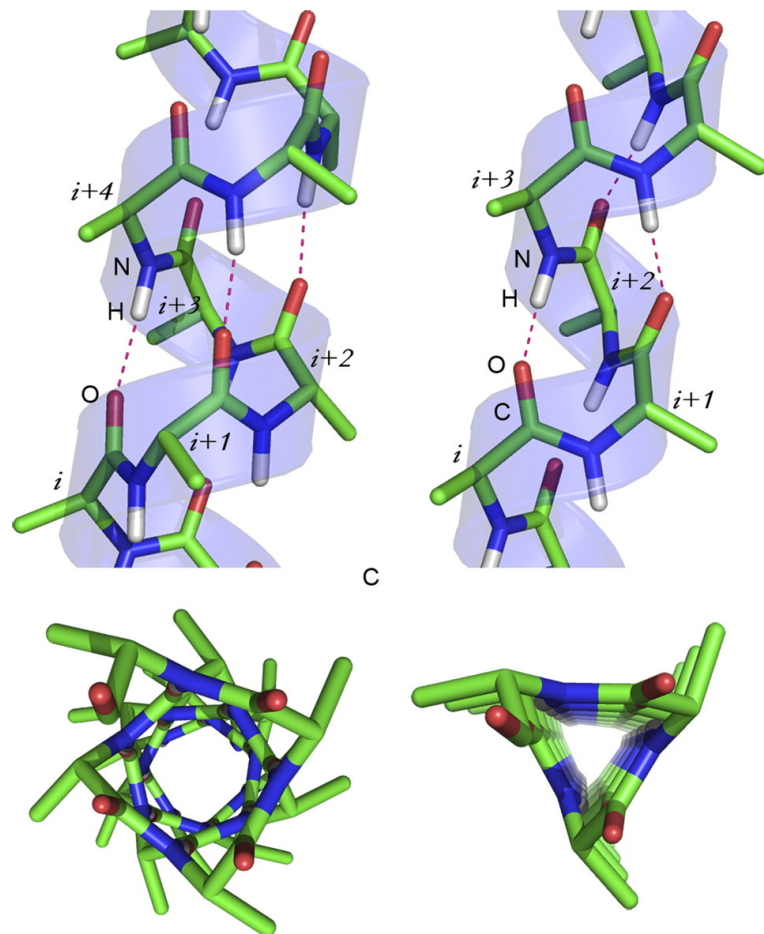


Unke et al., *Sci. Adv.* eadn4397 (2024)

Vieira-Pires et al., *J. Gen. Physiol.* 136, 585 (2010)

Bolin and Millhauser, *Acc. Chem. Res.* 32, 12, 1027 (1999)

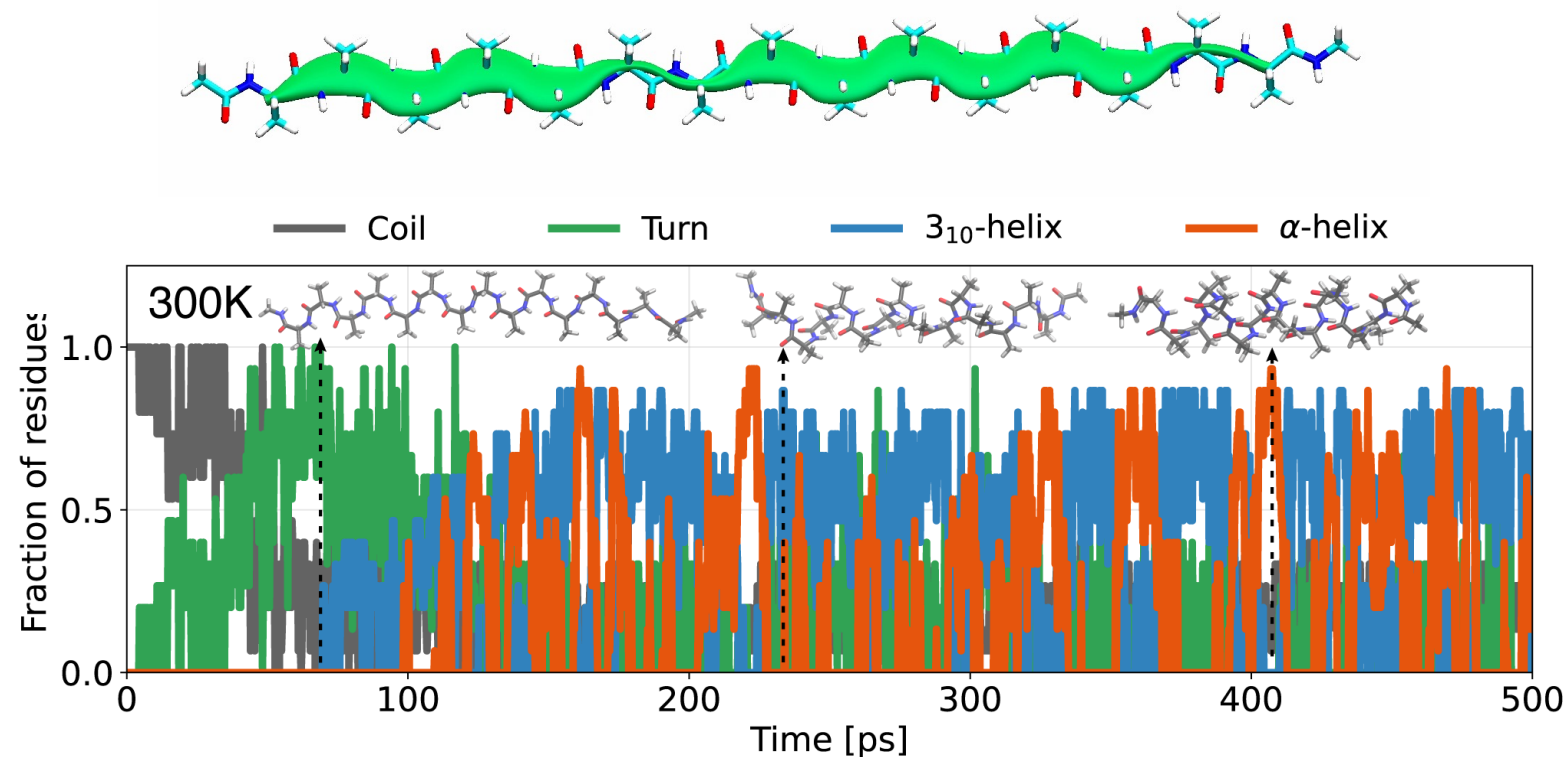
Folding of polyalanine



Canonical α -helix

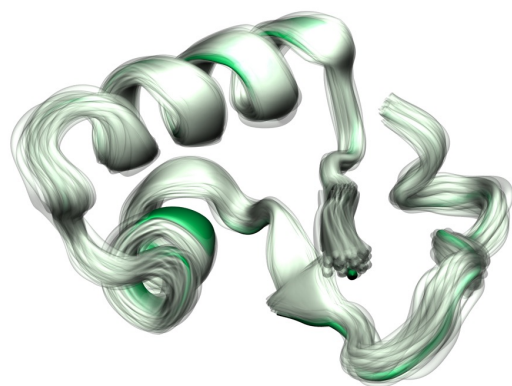
Canonical 3_{10} helix

AcAla₁₅NMe

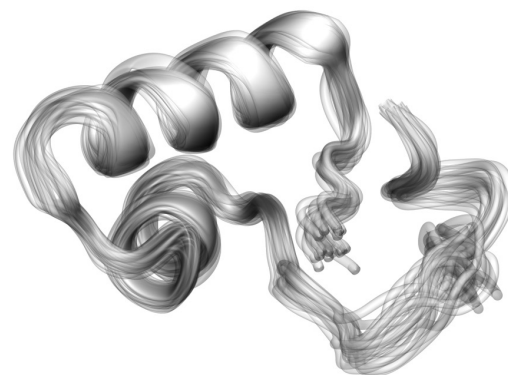


- Folding and alterations between alpha and 3_{10} forms
- Delicate interplay of hydrogen bonding, polarization, and dispersion interactions
- No system-specific Ala₁₅ data

Crambin

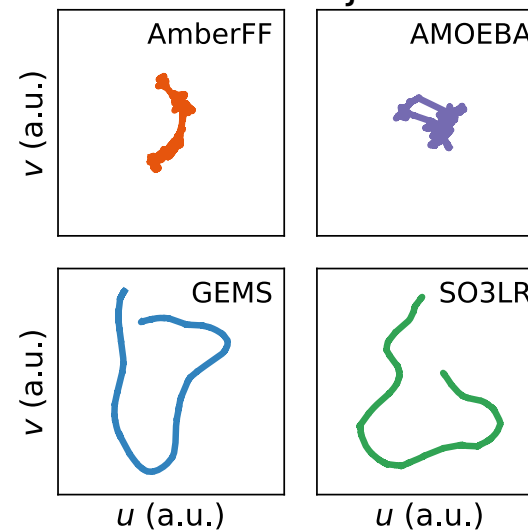


SO3LR MD Trajectory
25k atoms, 3 ns

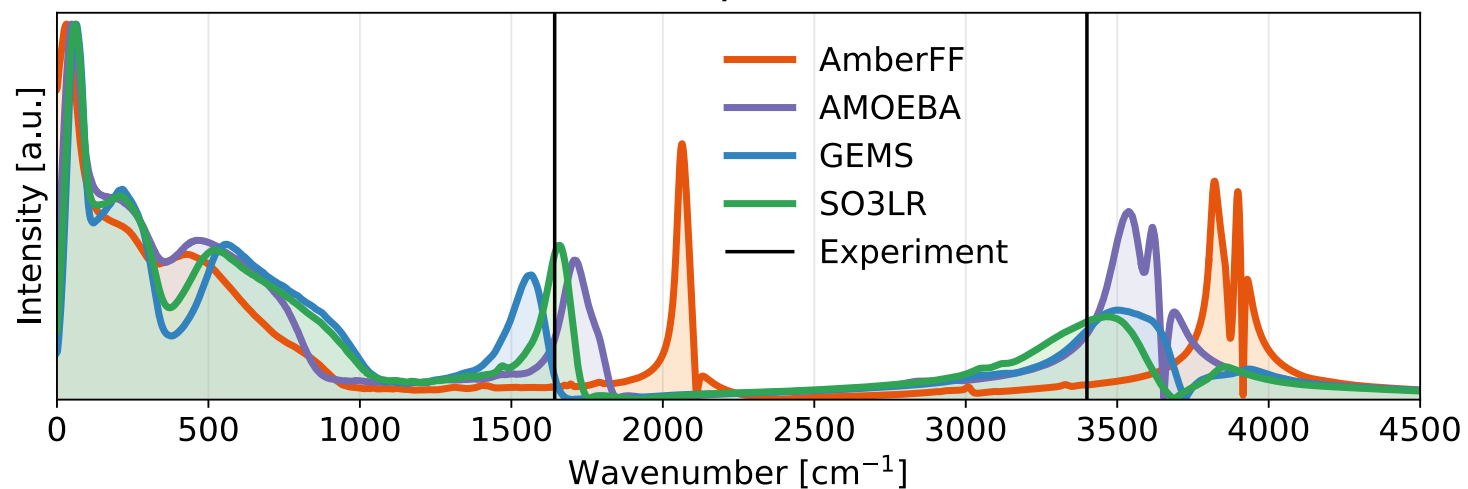


Experimental
structures

UMAP Projection



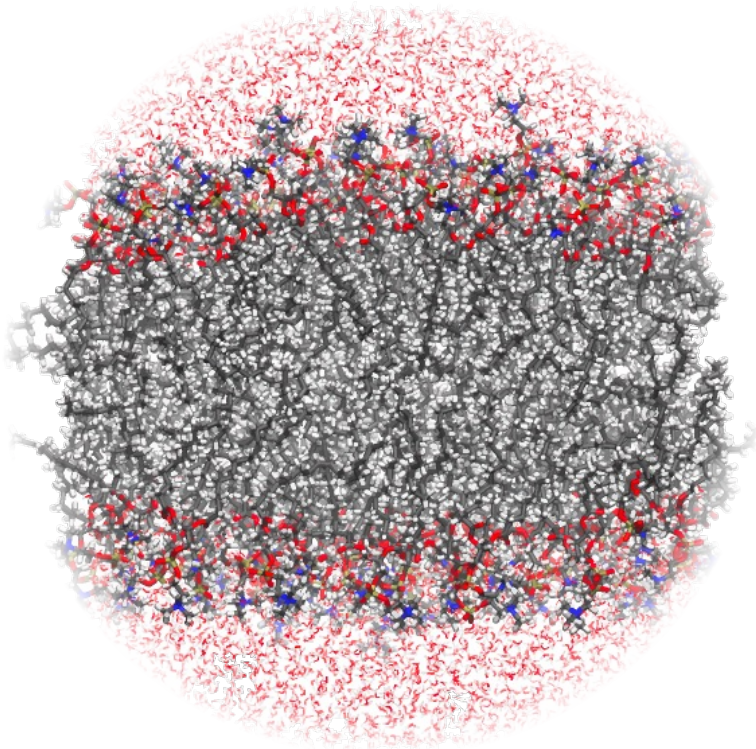
Power spectra



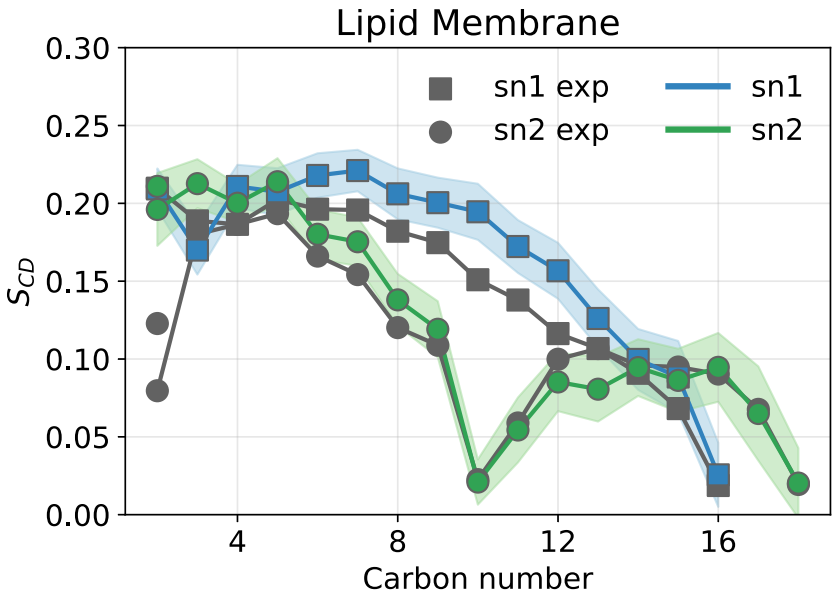
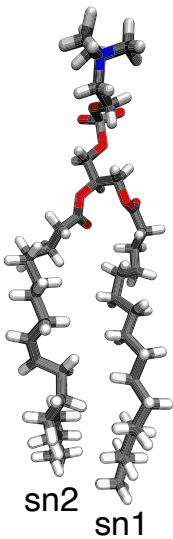
POPC Lipid bilayer

TABLE II. POPC Lipid Bilayer Structural Properties: bilayer thicknesses D_{HH} (Å) and area per lipid (Å²).

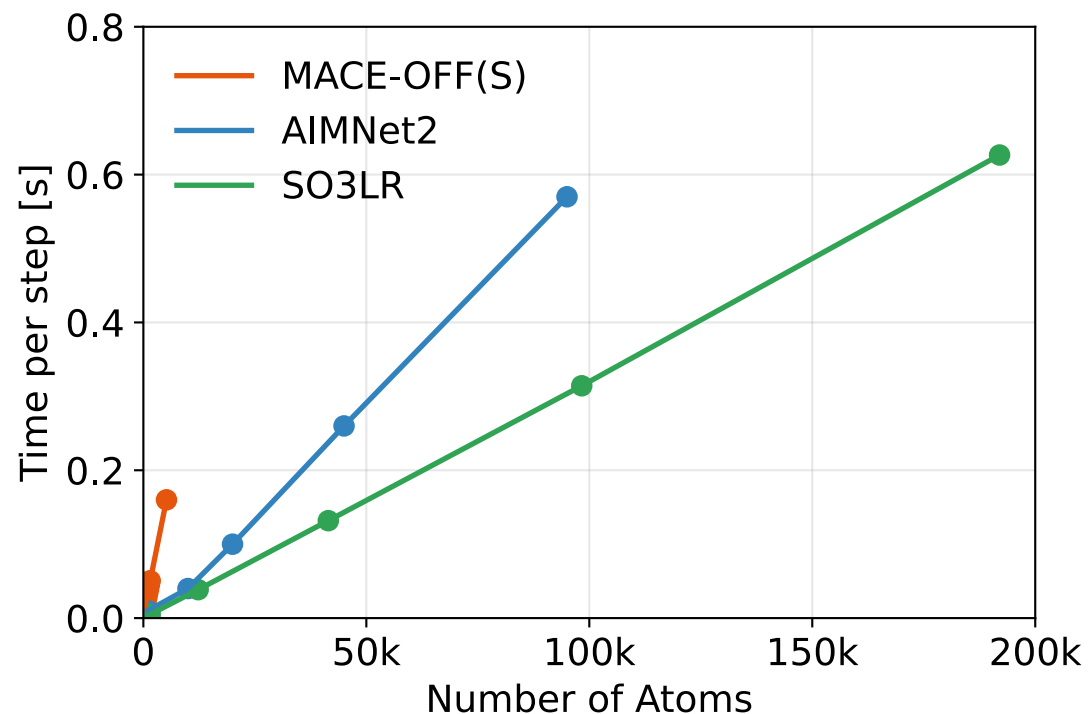
	D_{HH}	S/lipid
Experiment [76]	36.5	64.3±1.3
Lipid21 [77]	38.50±0.20	63.92±0.09
CHARMM36/LJ-PME [78]	37.3±0.30	65.4±0.5
SO3LR	37.0	58.0±0.1



33k atoms
500 ps



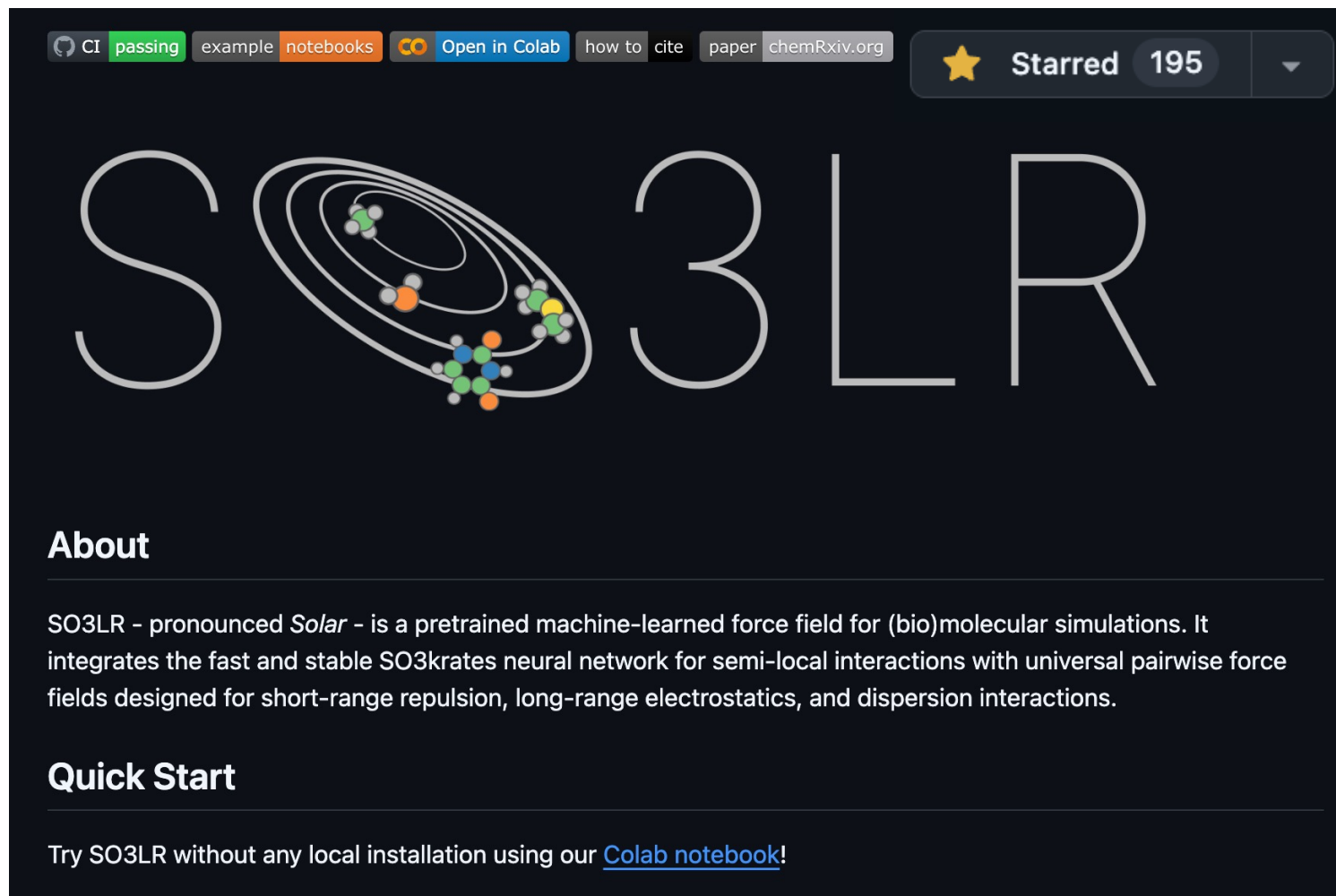
Single GPU performance



- $\sim 3 \mu\text{s}/\text{atom}/\text{step}$ on a single H100 GPU (2.6 ns/day for a 10,000-atom system)
- approaching sizes and timescales relevant for realistic biomolecules, while preserving QM accuracy

Model, Data, and Code

github.com/general-molecular-simulations/so3lr



CI passing example notebooks Open in Colab how to cite paper chemRxiv.org Starred 195

SO3LR

About

SO3LR - pronounced *Solar* - is a pretrained machine-learned force field for (bio)molecular simulations. It integrates the fast and stable SO3krates neural network for semi-local interactions with universal pairwise force fields designed for short-range repulsion, long-range electrostatics, and dispersion interactions.

Quick Start

Try SO3LR without any local installation using our [Colab notebook](#)!

- ✓ Works with ASE/JAX-MD
- ✓ Google Colab/Tutorials
- ✓ Command Line Interface
`so3lr nvt --input mol.xyz`



Outline

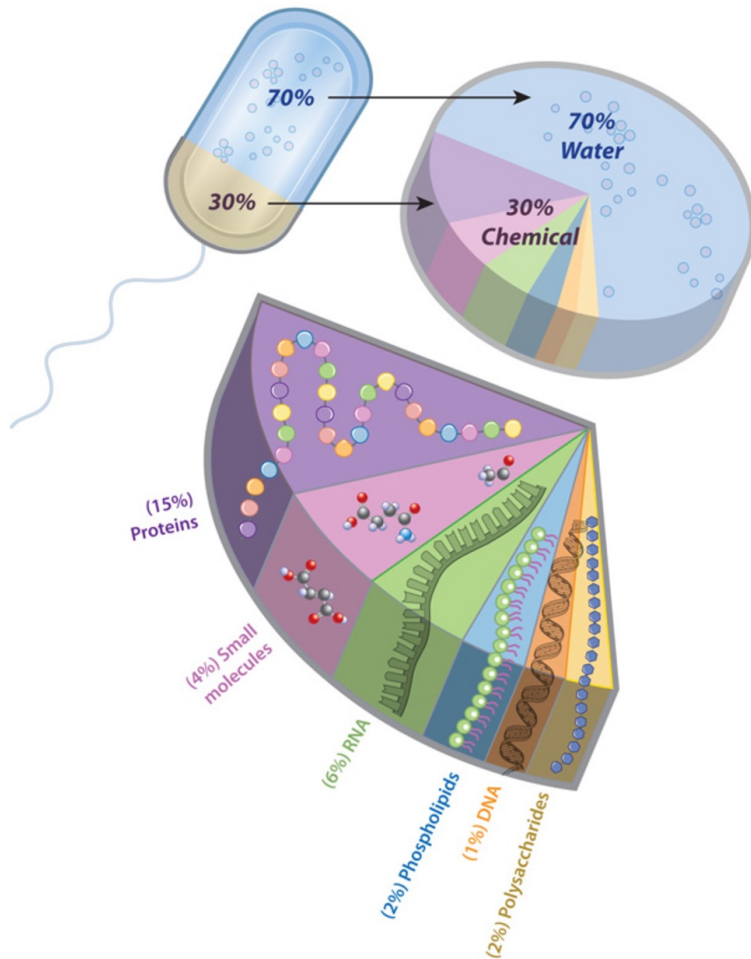
- *Introduction*
- *SO3LR: General-purpose MLFF for organic matter*
- ***QCell: QM dataset spanning diverse biomolecular fragments***
- *Conclusions & Outlook*

A. Kabylda, S. Suárez-Dou, Nils Davoine, Florian N. Brunig, A. Tkatchenko

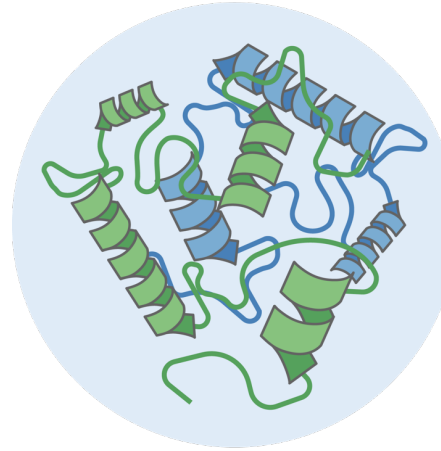
QCell: Comprehensive Quantum-Mechanical Dataset Spanning Diverse Biomolecular Fragments

AI Sci. 2, 025003 (2026)

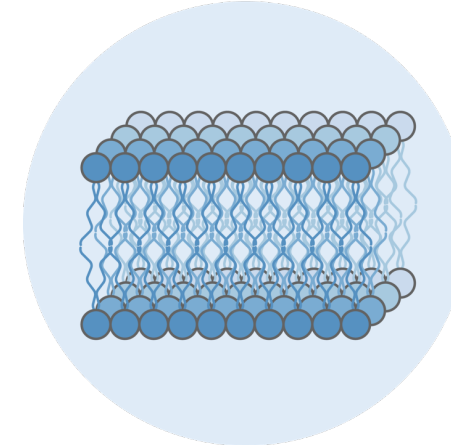
Major classes of biomolecules



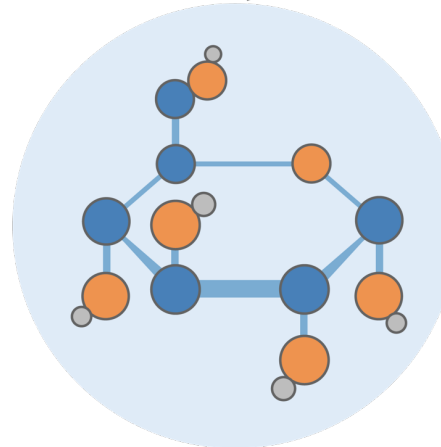
Proteins



Lipids



Carbohydrates



Nucleic Acids



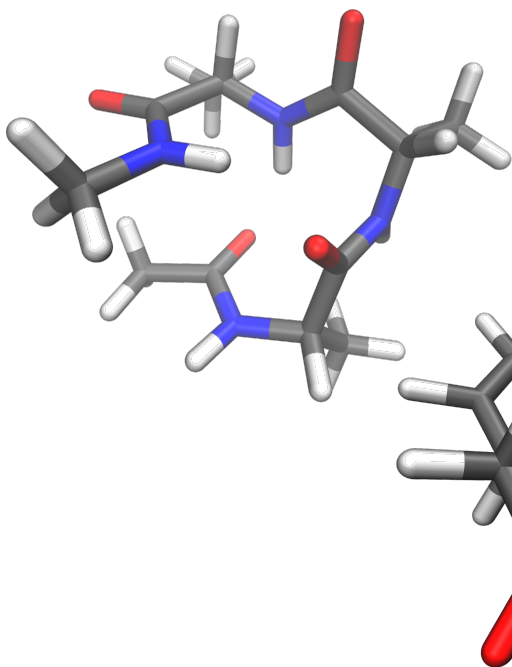
MD22 Dataset

Units of four major types of **biomolecules** and **supramolecules**
(10-90 ps dynamics - PBE+MBD level of theory at 500K):



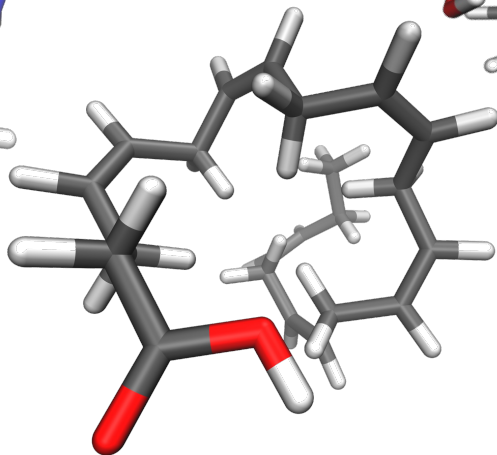
Proteins

Tetrapeptide
Ac-Ala3-NHMe



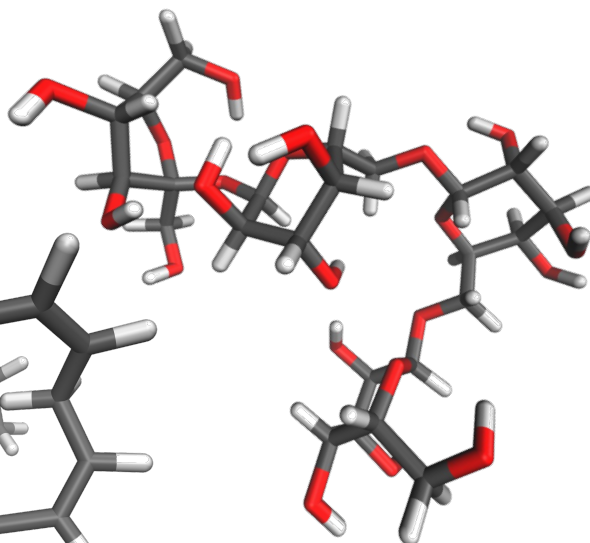
Lipids

Fatty acid
DHA



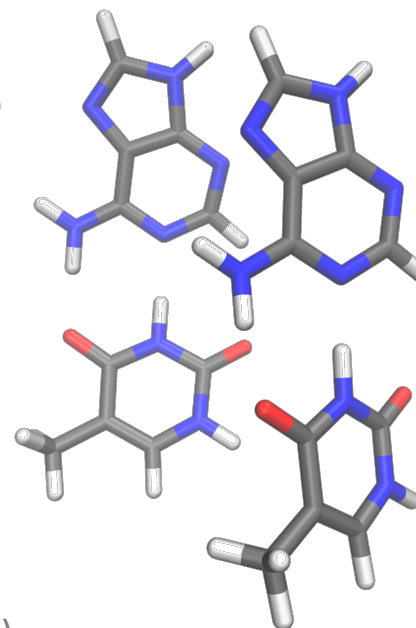
Carbohydrates

Tetrasaccharide
Stachyose



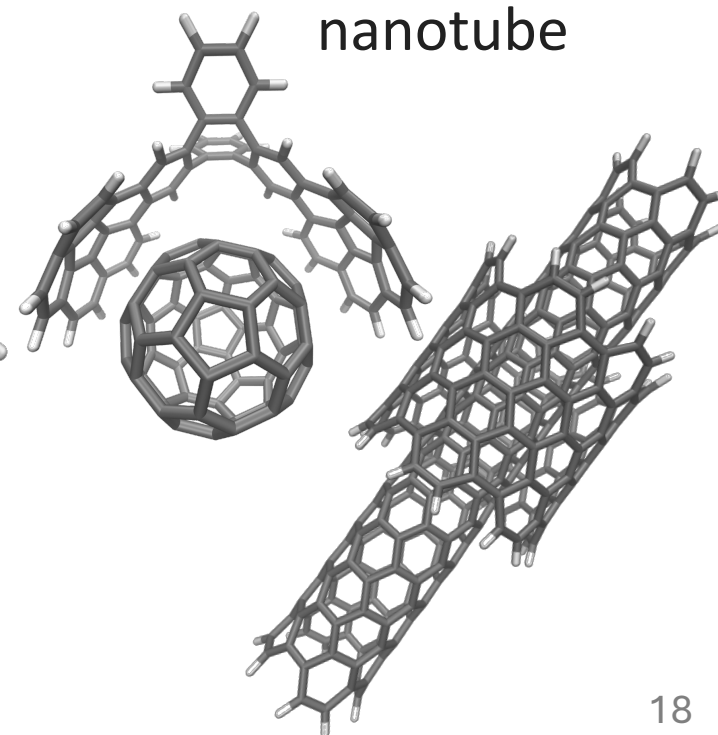
Nucleic acids

DNA base pairs
AT-AT-CG-CG
AT-AT



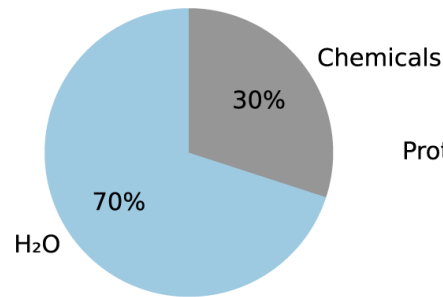
Supramolecules

Buckyball Catcher
Double-walled
nanotube

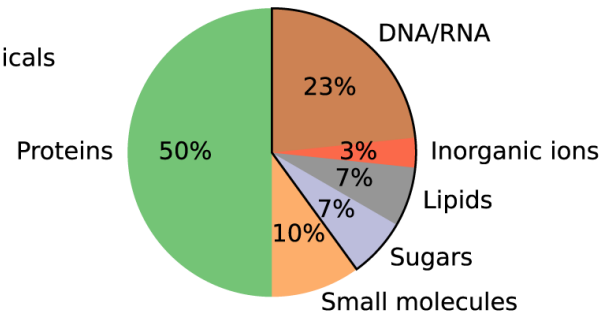


QCell: Biomolecular Building Blocks

A

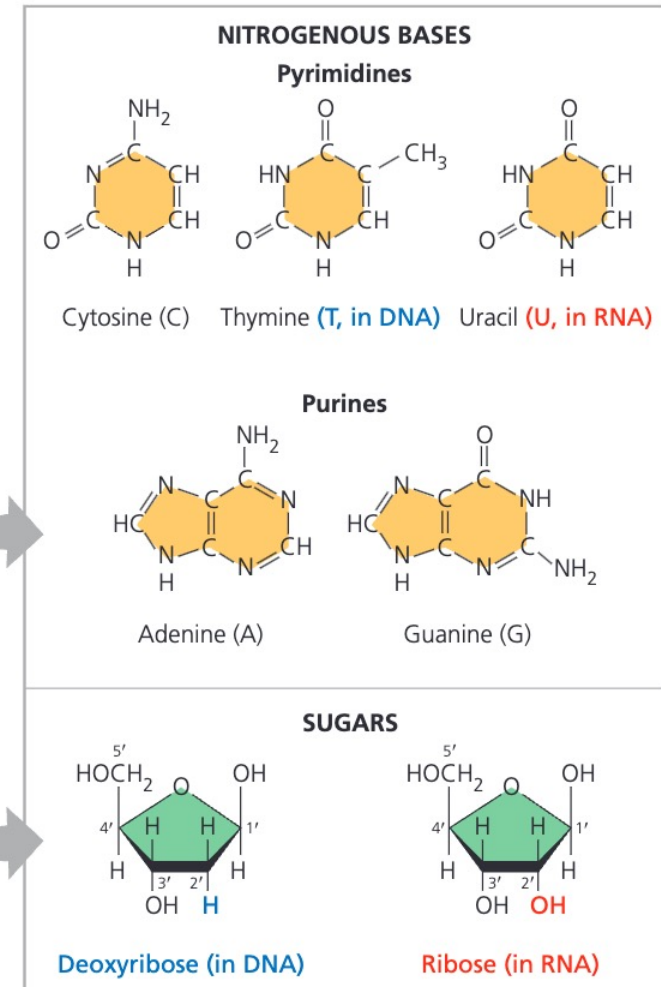
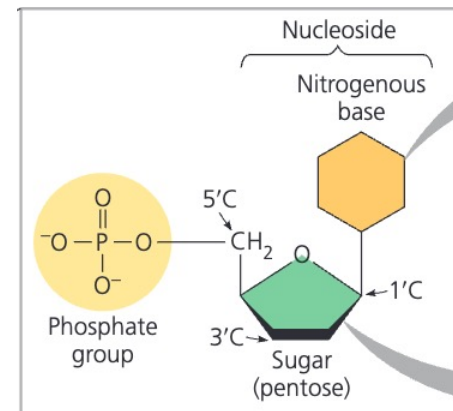
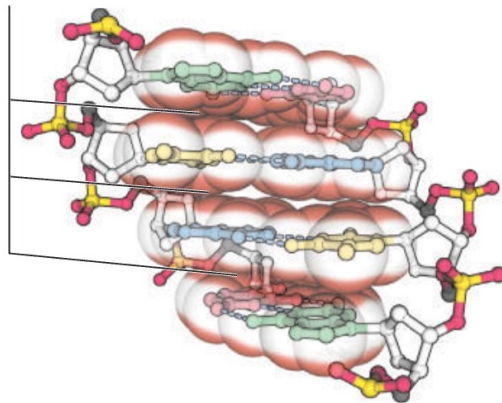


Bacterial Cell Composition



Breakdown of Chemicals (30%)

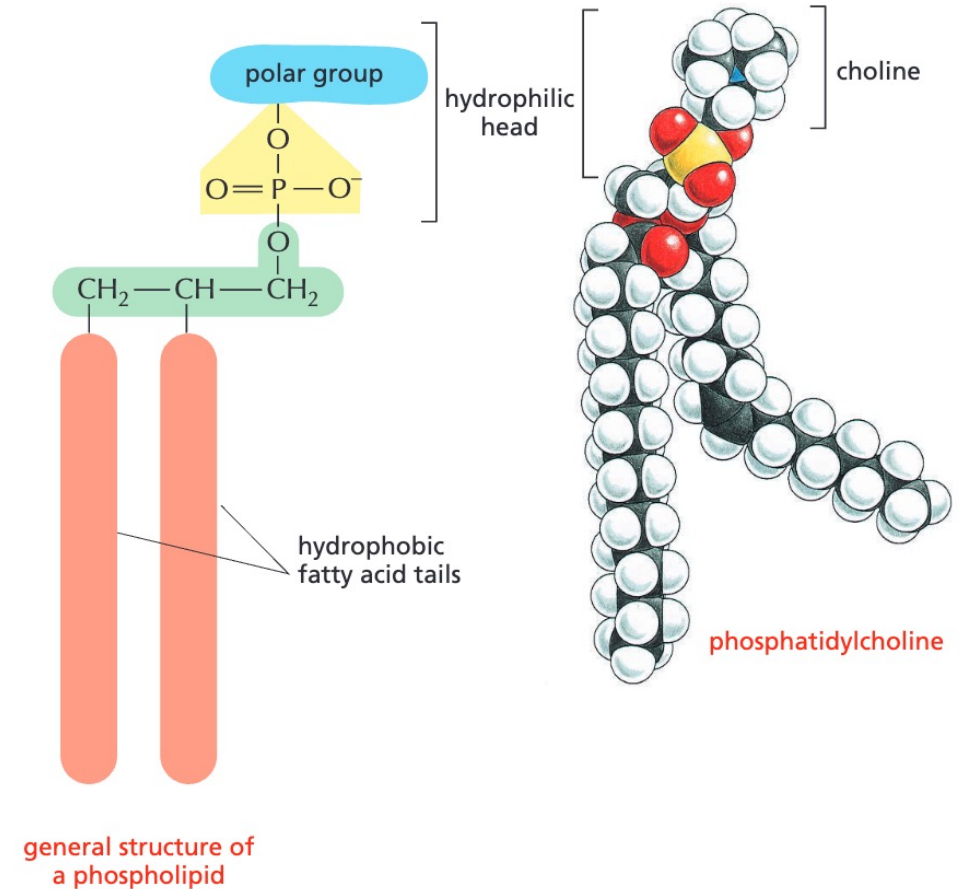
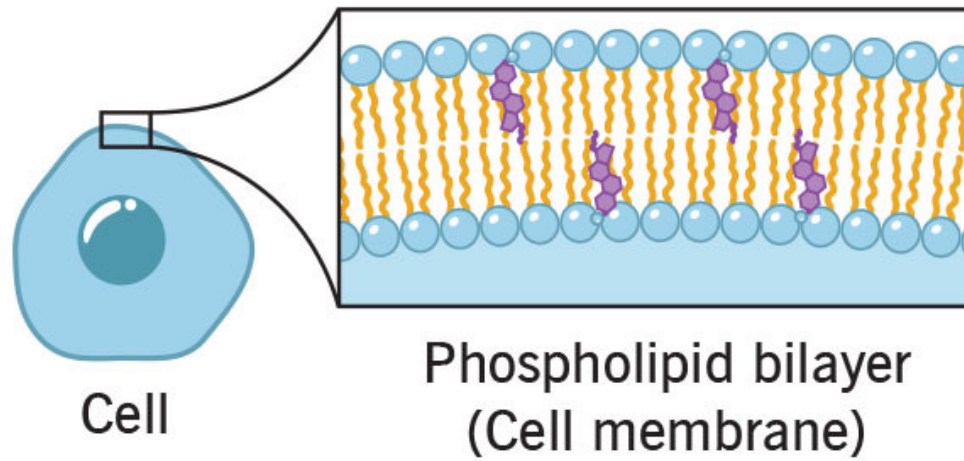
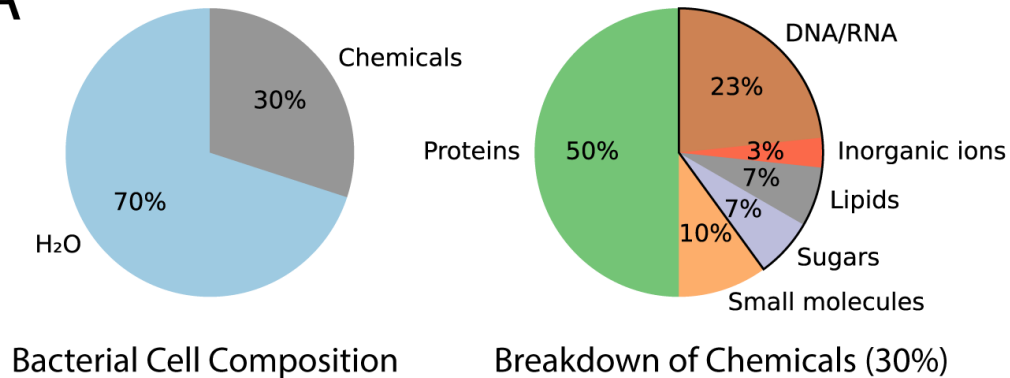
Base stacking
(van der Waal interactions)



- Gas-phase RNA fragments
- Solvated double-helical DNA dimers/trimers in canonical A, B, and Z forms

QCell: Biomolecular Building Blocks

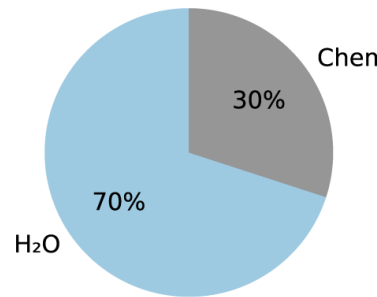
A



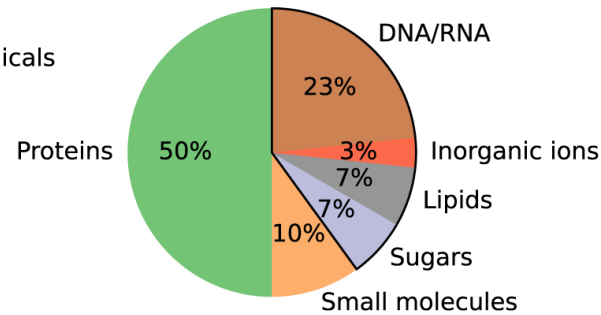
- POPC, POPE, POPG, and POPS 1-2-3-mers (with cholesterol)

QCell: Biomolecular Building Blocks

A



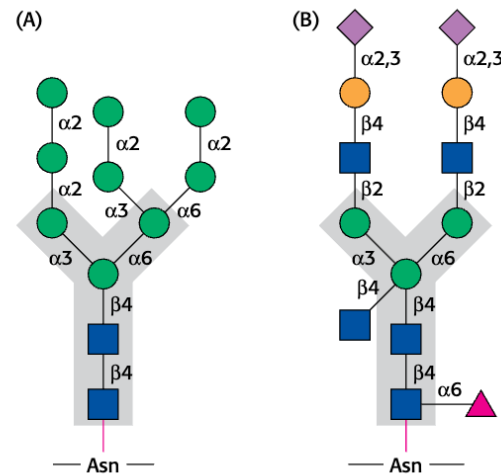
Bacterial Cell Composition



Breakdown of Chemicals (30%)

Abbreviations for sugars

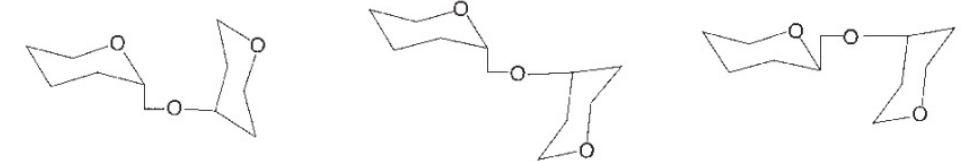
Fuc	▲	Fucose
Gal	●	Galactose
GalNAc	■	N-Acetylgalactosamine
Glc	●	Glucose
GlcNAc	■	N-Acetylglucosamine
Man	●	Mannose
Sia	◆	Sialic acid



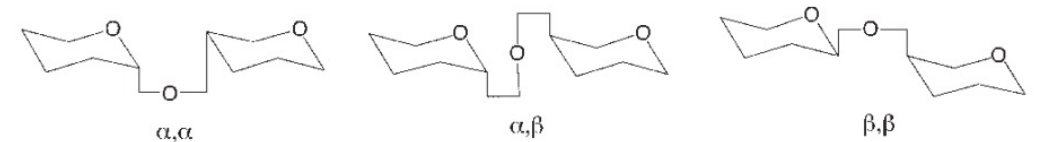
1,1-linkages (e.g. trehalose)



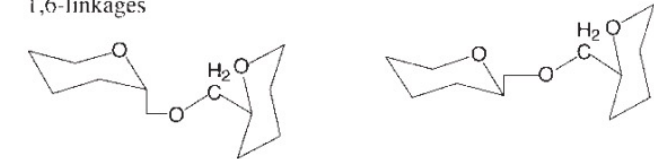
1,3-linkages (e.g. nigerose, laminarabiose)



1,4-linkages (e.g. galabiose, maltose, cellobiose)



1,6-linkages

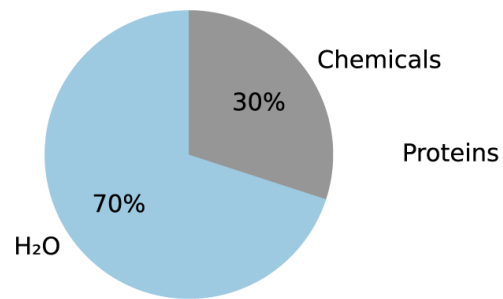


Example glycosyl linkages found in polysaccharides. Hydroxyl groups have been omitted for clarity.

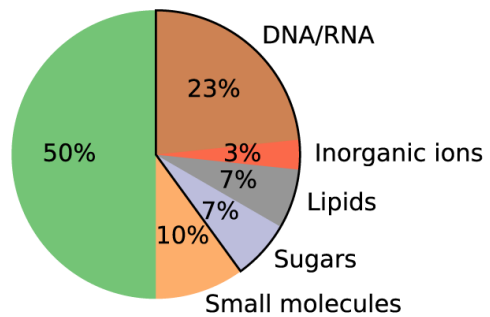
- 52 common monosaccharides, including both pentose and hexose structures in α and β anomeric configurations; glycosidic linkages

QCell: Biomolecular Building Blocks

A



Bacterial Cell Composition



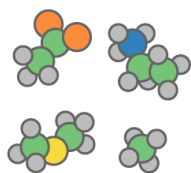
Breakdown of Chemicals (30%)

B

1. Selection of systems
2. Conformational sampling
3. Fragment selection
4. Pre-optimization: DFTB+MBD
5. QM calculations: PBE0+MBD(-NL)

C

Small Molecules



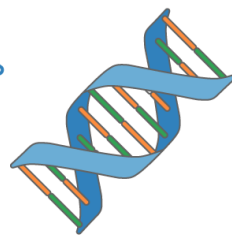
QM7X
QCML
AQM

Proteins



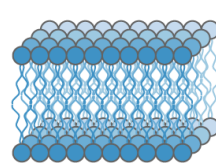
GEMS

Nucleic Acids



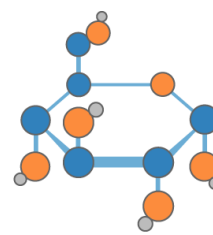
DNA/RNA
fragments

Lipids



Fatty Acid
fragments
(with cholesterol)

Sugars



Disaccharides
Glycosidic
linkages

Ions



Solvated
ions

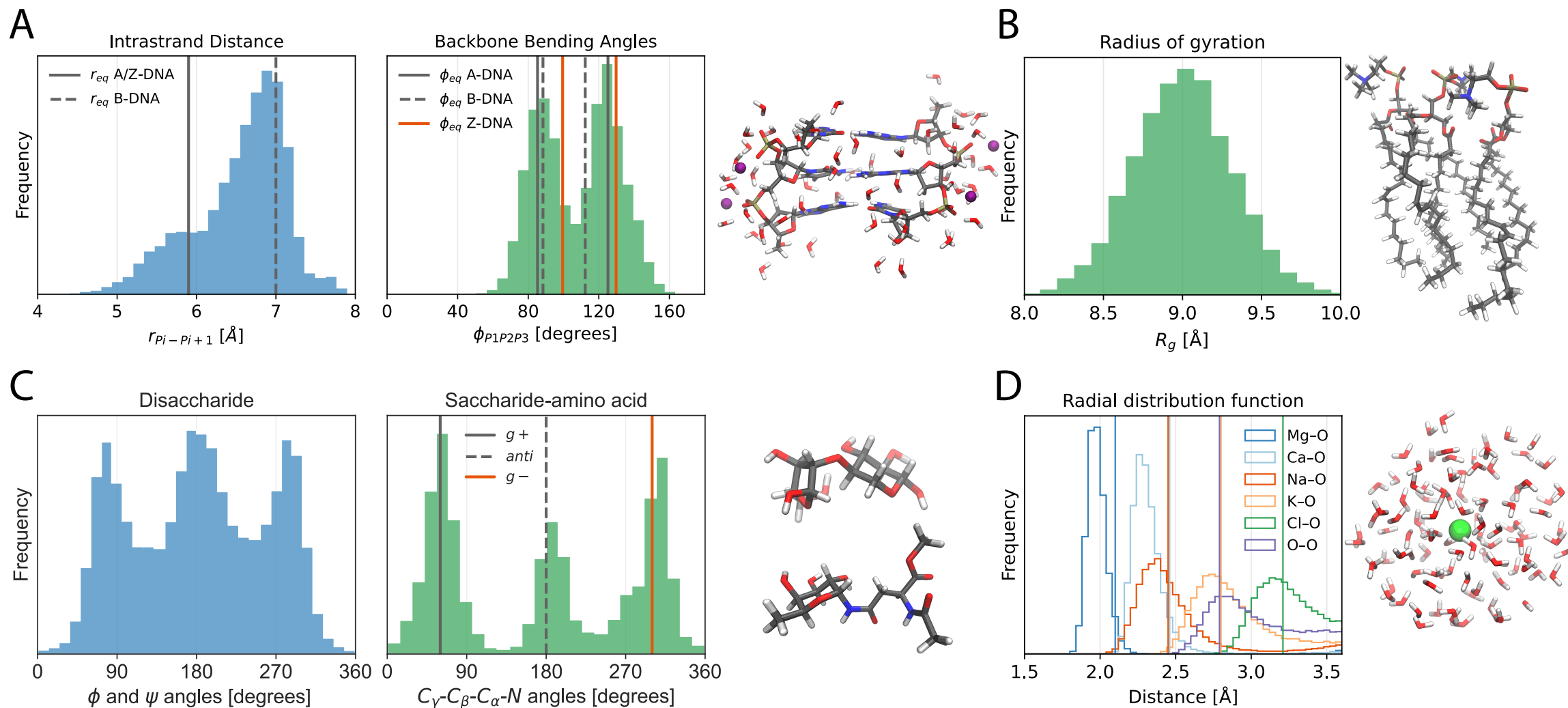
Dimers



DES370k
dimers

Newly generated data

QCell: QM Dataset Spanning Diverse Biomolecular Fragments



Kabylda et al., *AI Sci* 2, 025003 (2026)

Alberts et al., *Molecular Biology of the Cell*

Acknowledgements



Alexandre
Tkatchenko



Sergio Suárez-
Dou



Almaz
Khabibrakhmanov



Leonardo M.
Sandonas



Nils
Davoine



Florian N.
Brünig



Valentin
Vassilev-Galindo



Igor
Poltavsky



Klaus-Robert
Müller



J. Thorben
Frank



Stefan
Chmiela



Oliver T.
Unke



Outline

- *Introduction*
- *SO3LR: General-purpose MLFF for organic matter*
- *QCell: QM dataset spanning diverse biomolecular fragments*
- **Conclusions & Outlook**

Conclusions & Outlook

General-Purpose MLFFs for (Bio)Molecular Simulations

- ✓ **SO3LR**: Pretrained MLFF integrating neural network with explicit long-range electrostatics and dispersion
- ✓ **QCell**: Quantum-mechanical data for diverse large (bio)molecular fragments
- ✓ **SO3LR v2**: Better, faster & stronger

Outlook

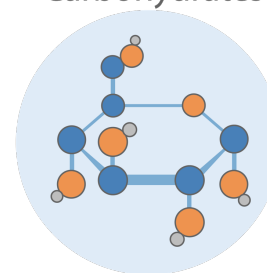
- ❑ Materials-Molecules Interfaces
- ❑ Extending simulations to treat NQE
- ❑ Multiscale coarse-grained MLFFs



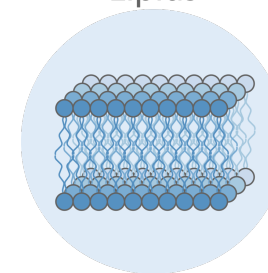
Proteins



Carbohydrates



Lipids



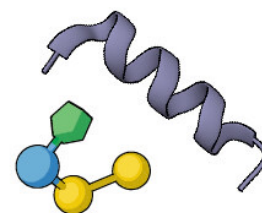
Nucleic Acids



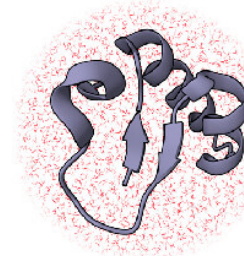
SO3LR

Pretrained Machine Learned Force Field for Molecular Simulations

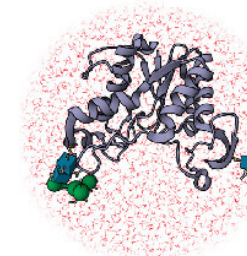
Small biomolecules



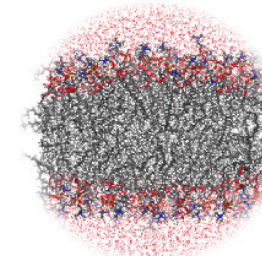
Protein



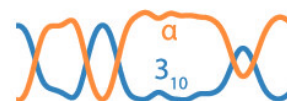
Glycoprotein



Lipid bilayer



Secondary structure



IR Spectra



RMSD

Carbohydrate
Protein



Electronic density



Kabylda et al., *JACS* 147, 37, 33723 (2025)

Kabylda et al., *AI Sci* 2, 025003 (2026)

Kabylda et al., *Nat. Commun.* 14, 3562 (2023)

Kabylda et al., *Adv. Funct. Mater.* 35, 2417891 (2025)

Esders, Schnake, Lederer, Kabylda et al., *JCTC* 21, 714 (2025)

Chmiela, Vassilev-Galindo, Unke, Kabylda et al., *Sci. Adv.* eadf0873 (2023)

