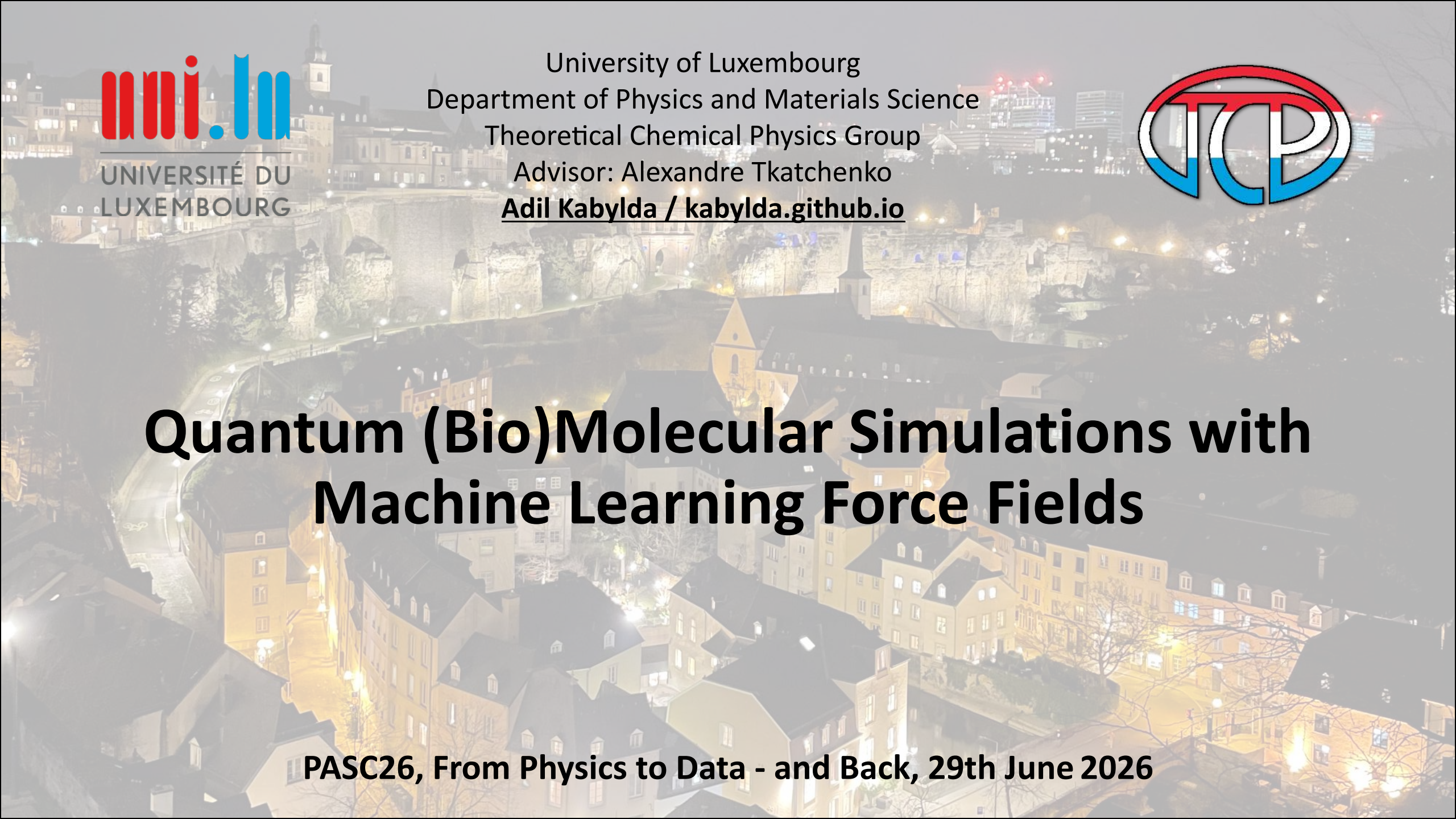


University of Luxembourg
Department of Physics and Materials Science
Theoretical Chemical Physics Group
Advisor: Alexandre Tkatchenko
Adil Kabylda / kabylda.github.io

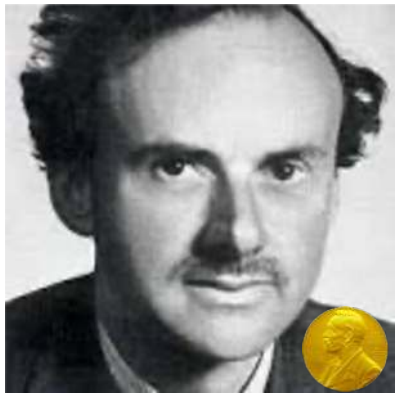
Quantum (Bio)Molecular Simulations with Machine Learning Force Fields

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



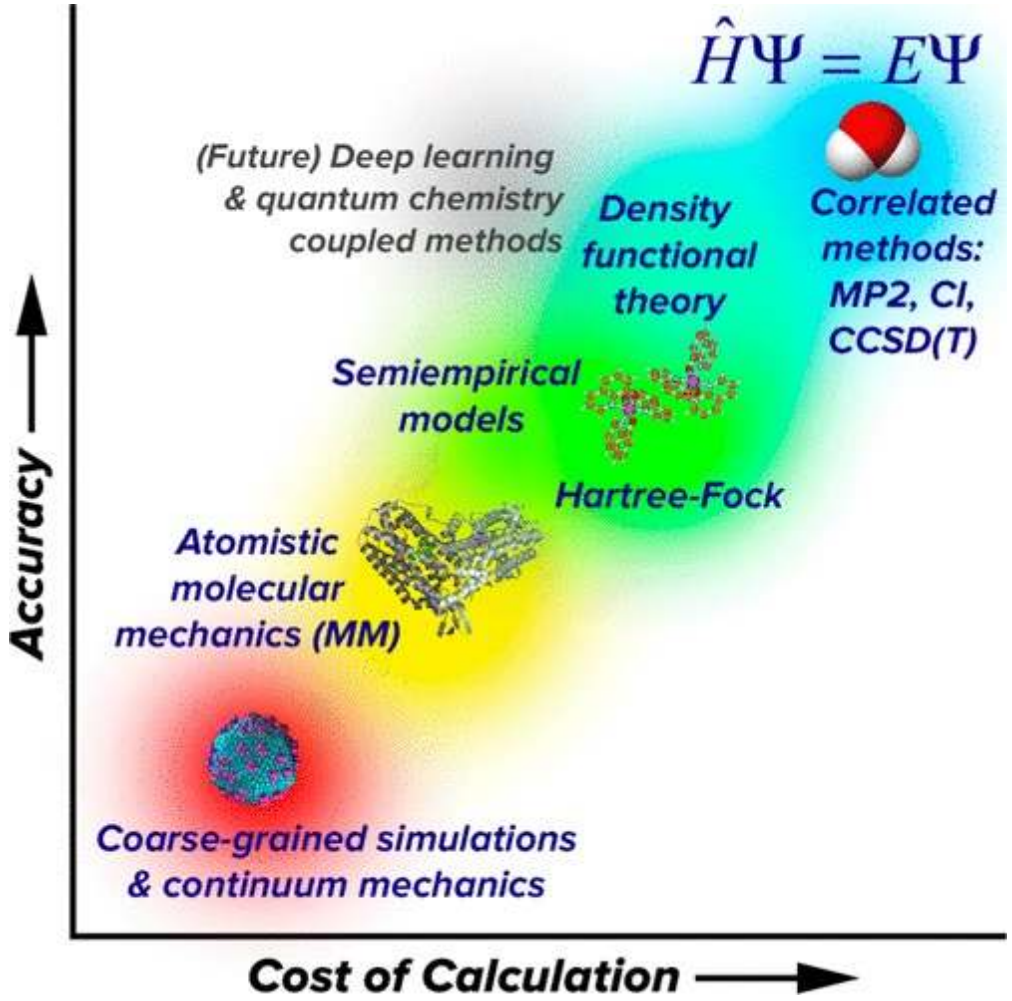
P. A. M. Dirac
Quantum mechanics of many-electron Systems
Proc. R. Soc. Lond., 123, 714 (1929)

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

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

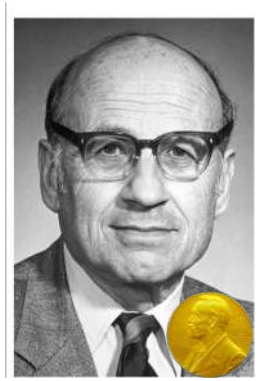
- Realizing Schrödinger's dream requires development of approximate practical methods

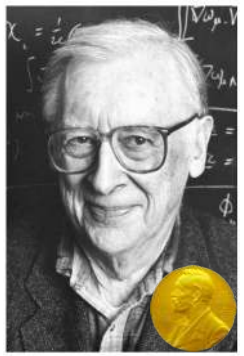
Hierarchy of Modelling Methods

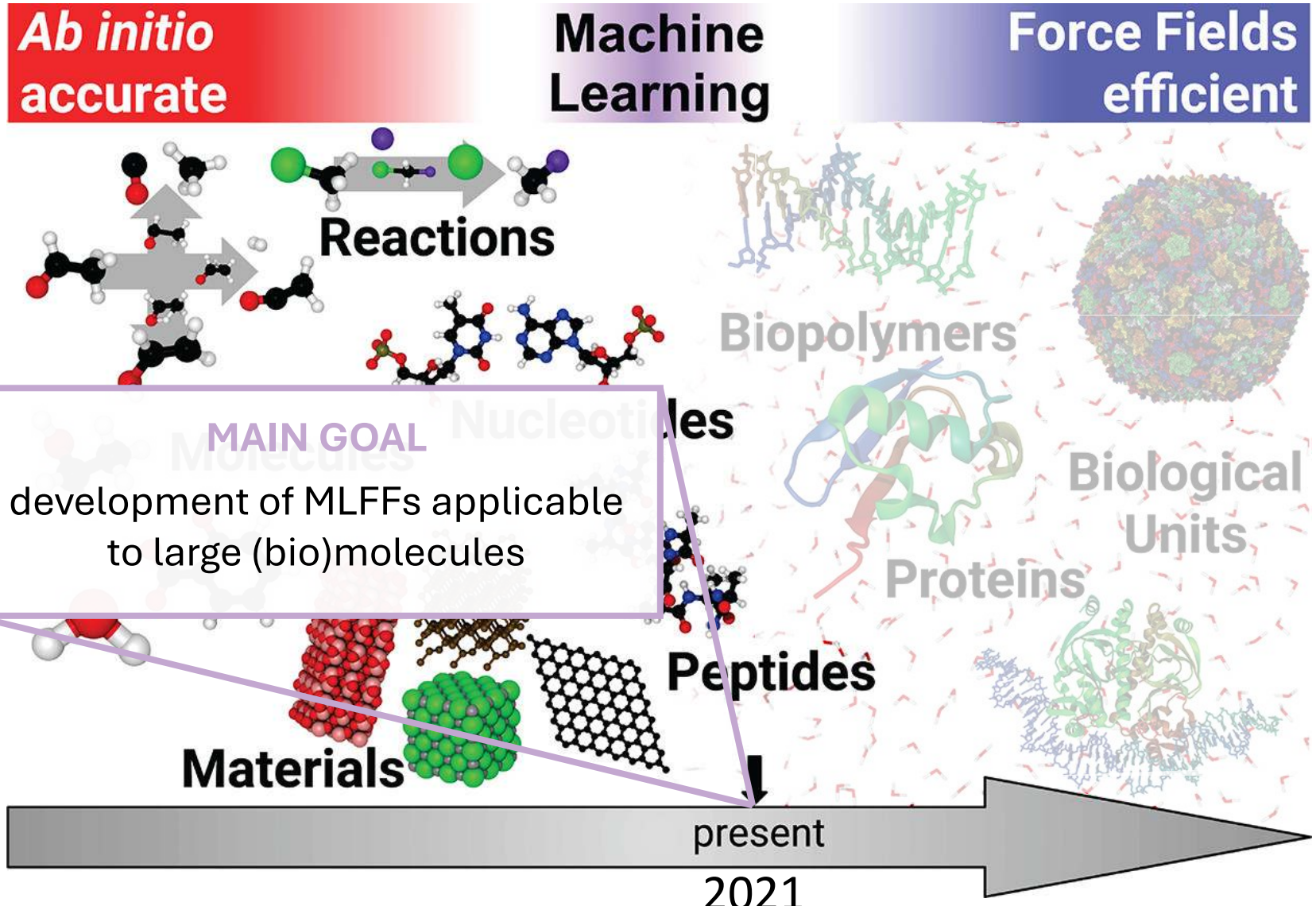


Borges et al., *Rev. Chem.* 121, 10, 5633 (2023)

Machine Learning Force Fields (MLFFs)

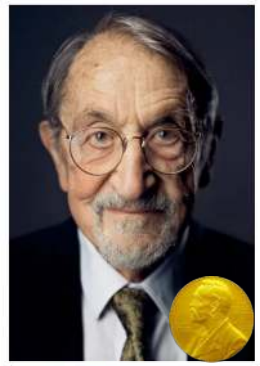


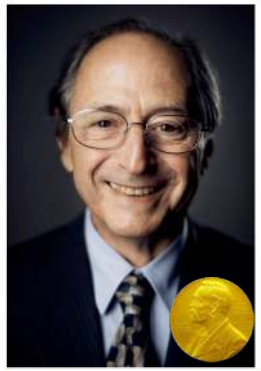


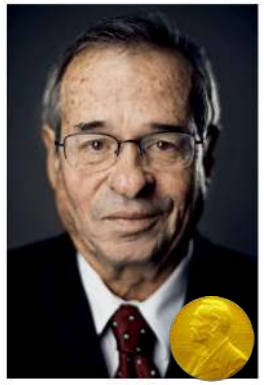


MAIN GOAL

development of MLFFs applicable to large (bio)molecules



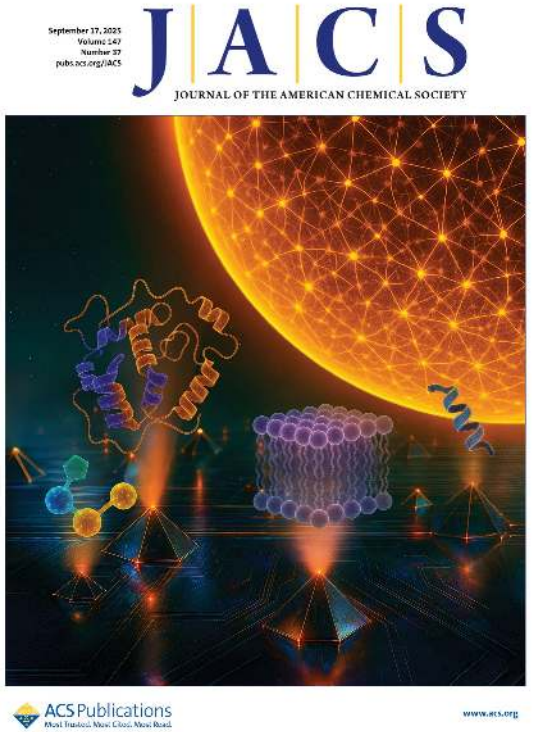




Unke et al., *Chem. Rev.* 121, 16, 10142 (2021)

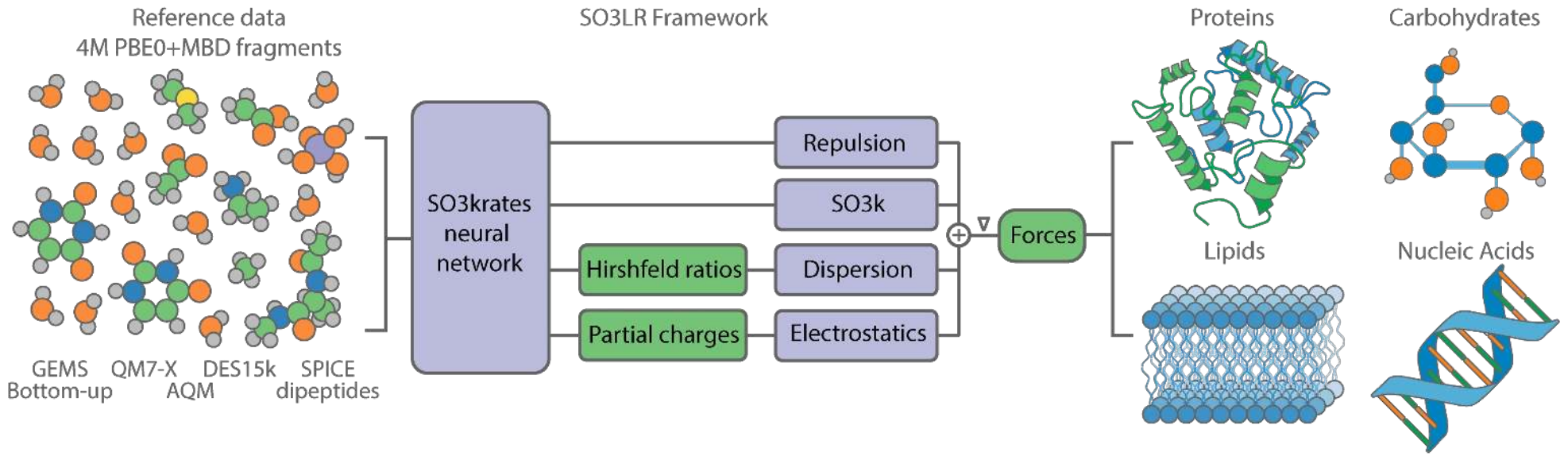
Outline

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



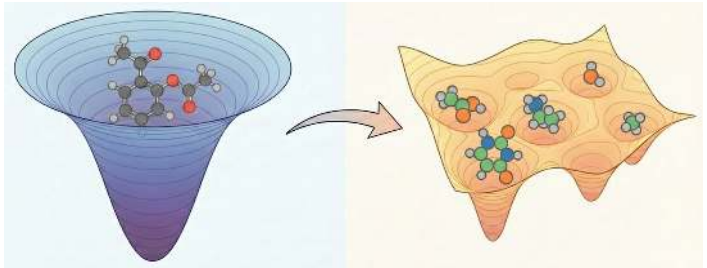
A. Kabylda, J. T. Frank, S. Suárez-Dou, A. Khabibrakhmanov, L. Medrano Sandonas, O. T. Unke, S. Chmiela, K.-R. Müller, A. Tkatchenko
Molecular Simulations with a Pretrained Neural Network and Universal Pairwise Force Fields
Journal of the American Chemical Society 147, 37, 33723 (2025). Cover article.

SO3LR: General-purpose MLFF for organic matter



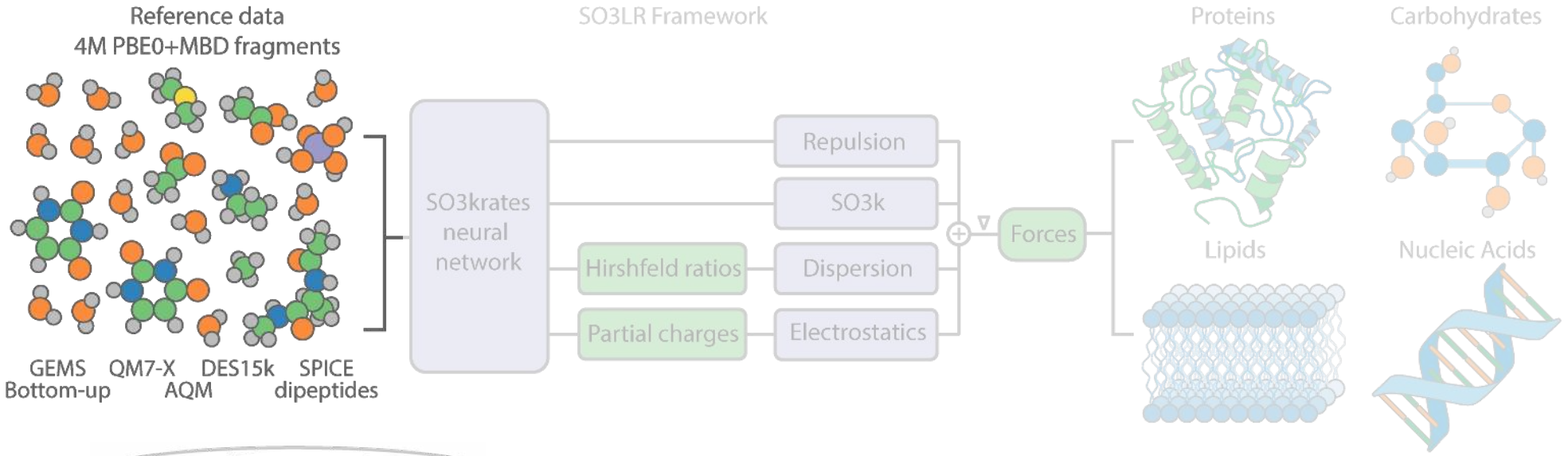
- MLFF architectures
- Quantum-mechanical datasets
- Physical interaction models
- GPU software/hardware

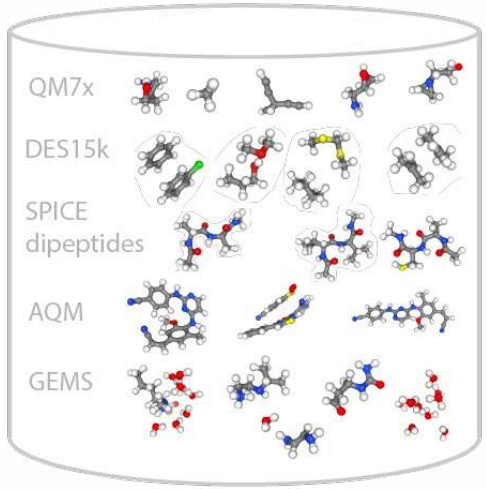
System-specific -> General-Purpose MLFFs

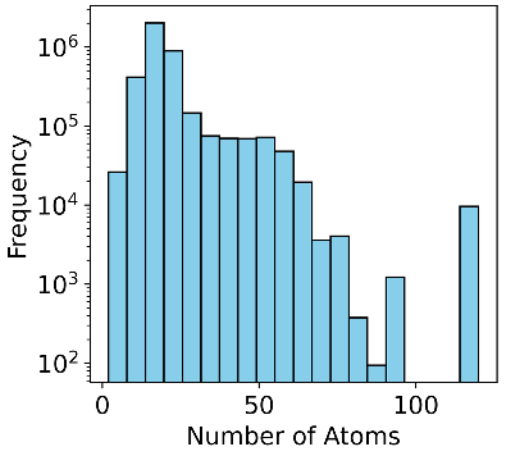


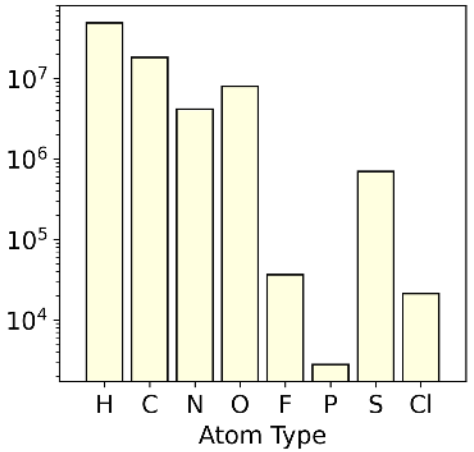
Kabylda et al., *JACS, 147, 37, 33723* (2025)

Diverse QM training dataset





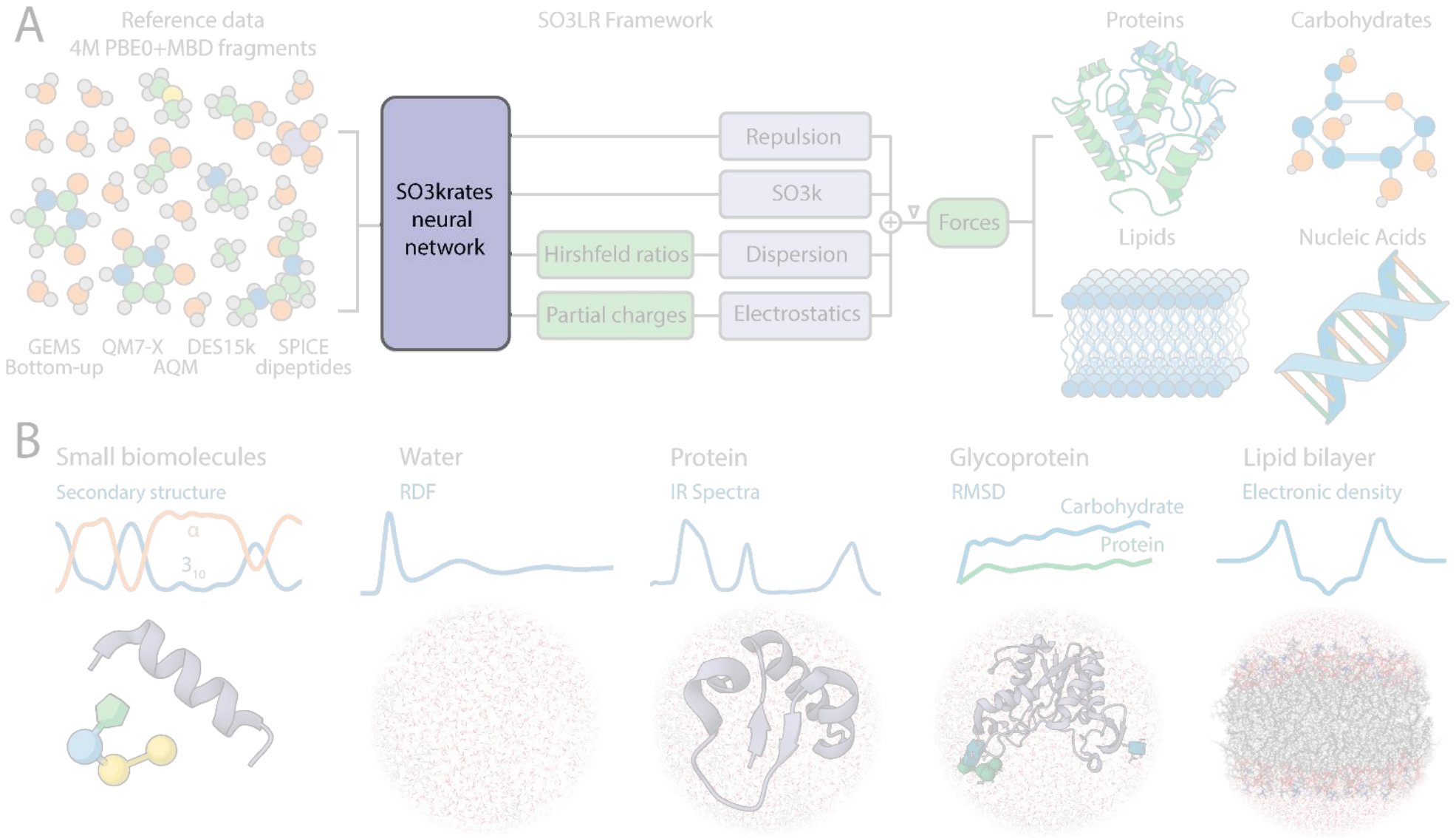




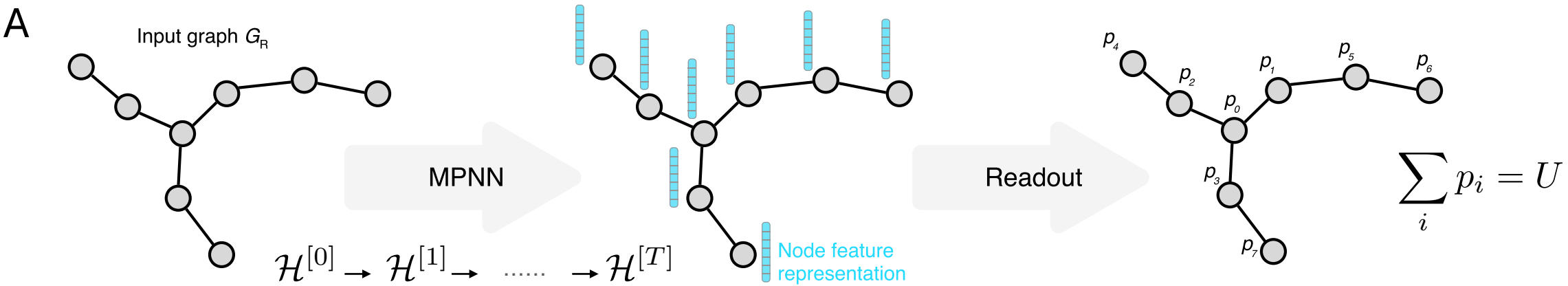
- PBE0+MBD: 3.86M structures

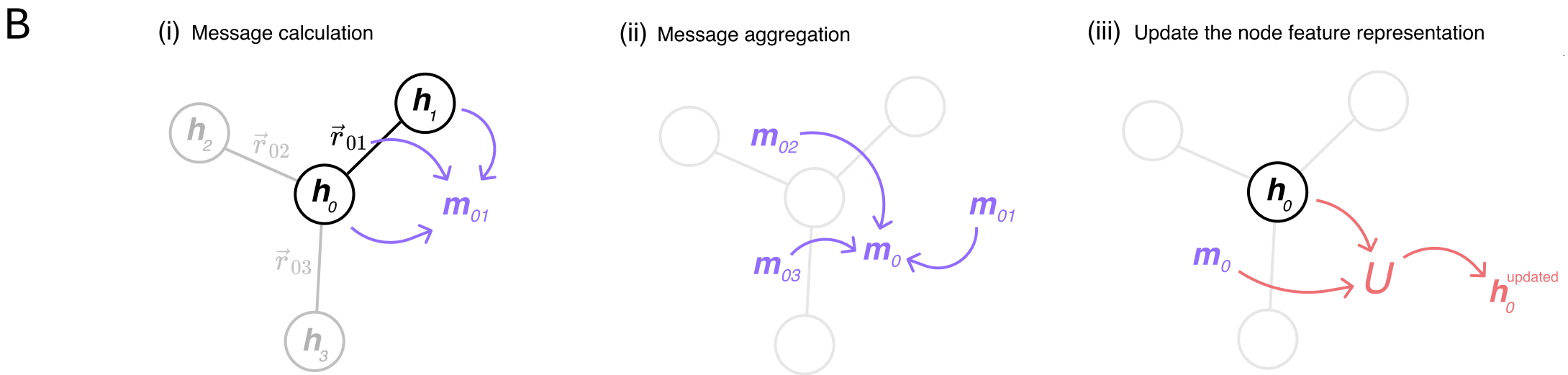
Kabylda et al., *JACS, 147, 37, 33723* (2025)

SO3LR

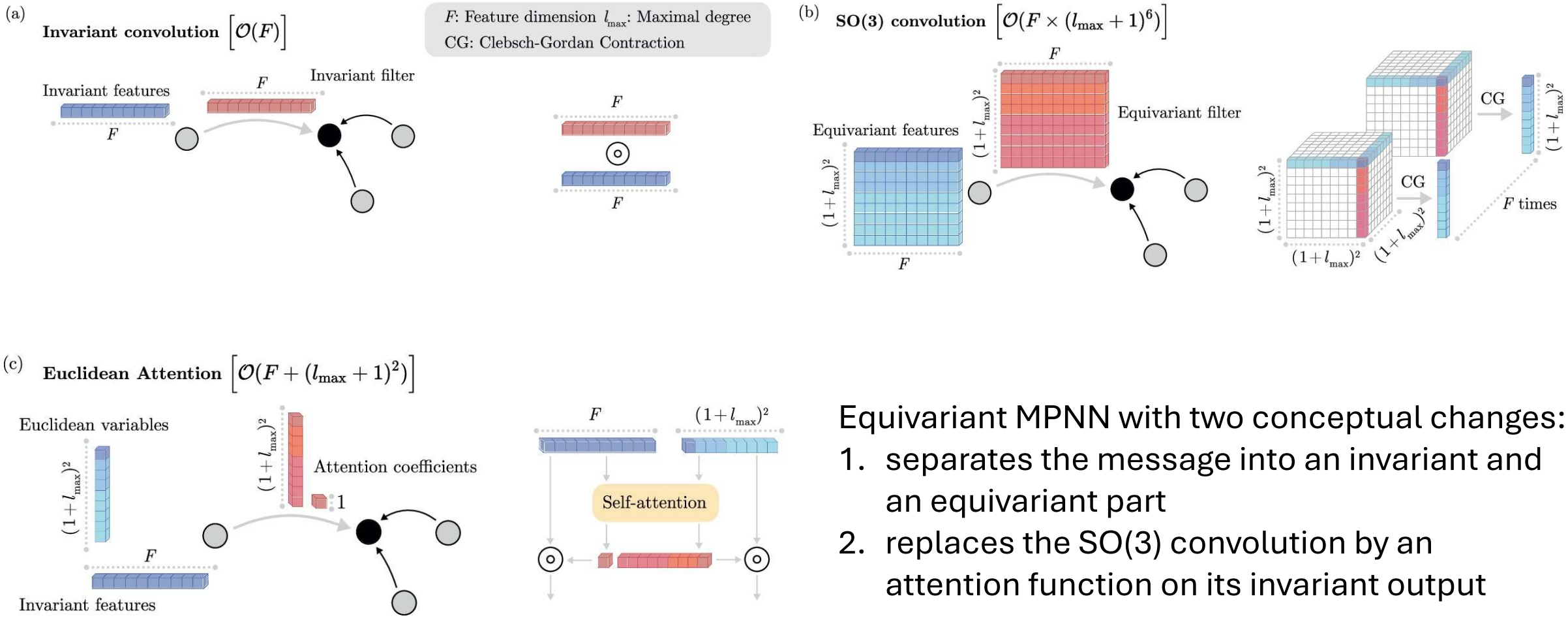

A
Reference data
4M PBE0+MBD fragments
GEMS
Bottom-up
QM7-X
AQM
DES15k
SPICE
dipeptides
SO3LR Framework
SO3krates neural network
Repulsion
SO3k
Hirshfeld ratios
Dispersion
Partial charges
Electrostatics
Forces
Proteins
Carbohydrates
Lipids
Nucleic Acids
B
Small biomolecules
Secondary structure
α
3₁₀
Water
RDF
Protein
IR Spectra
Glycoprotein
RMSD
Carbohydrate
Protein
Lipid bilayer
Electronic density

Message Passing Neural Networks





SO3krates – cornerstone of SO3LR

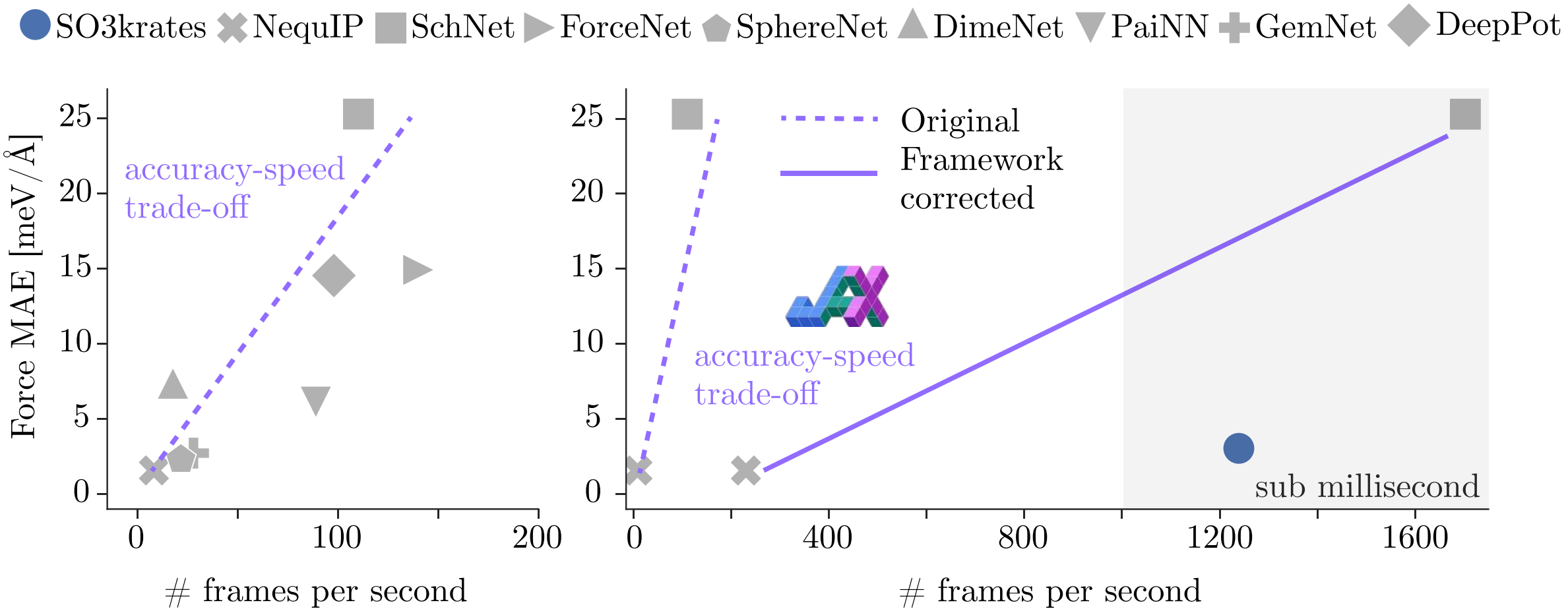


Equivariant MPNN with two conceptual changes:

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

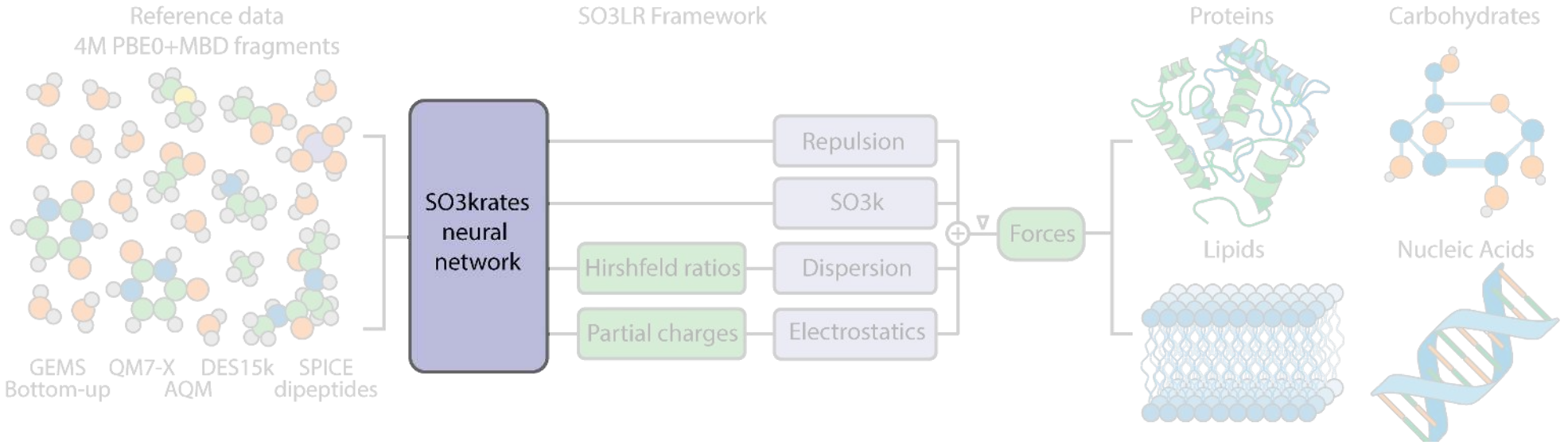
Frank et al., *Nat. Commun.* 15, 6539 (2024)

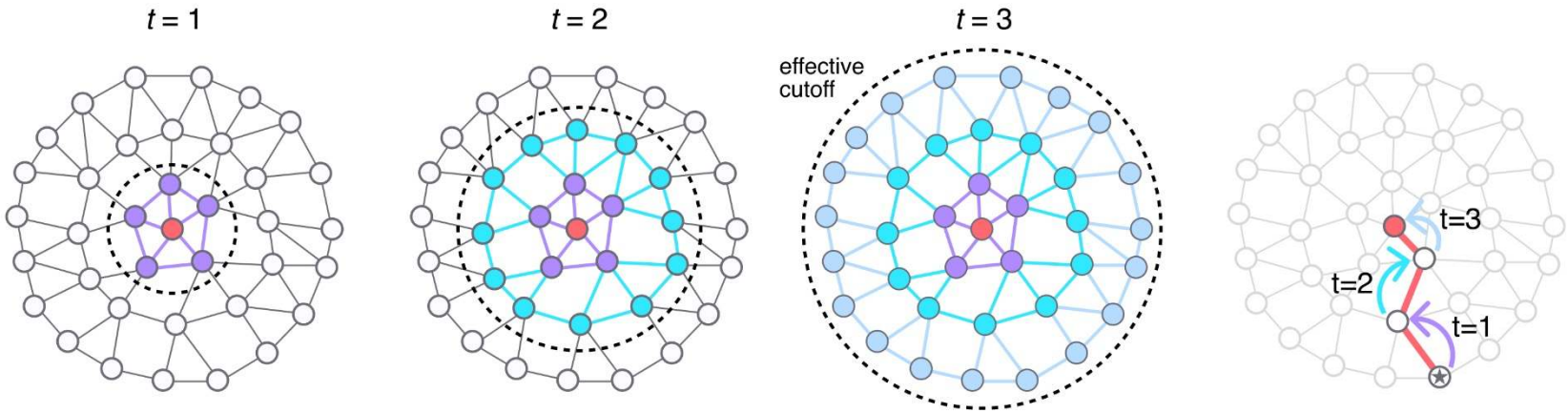
MLFFs accuracy-speed trade-off



Fu et al., *TMLR* (2022)
Frank et al., *Nat. Commun.* 15, 6539 (2024)

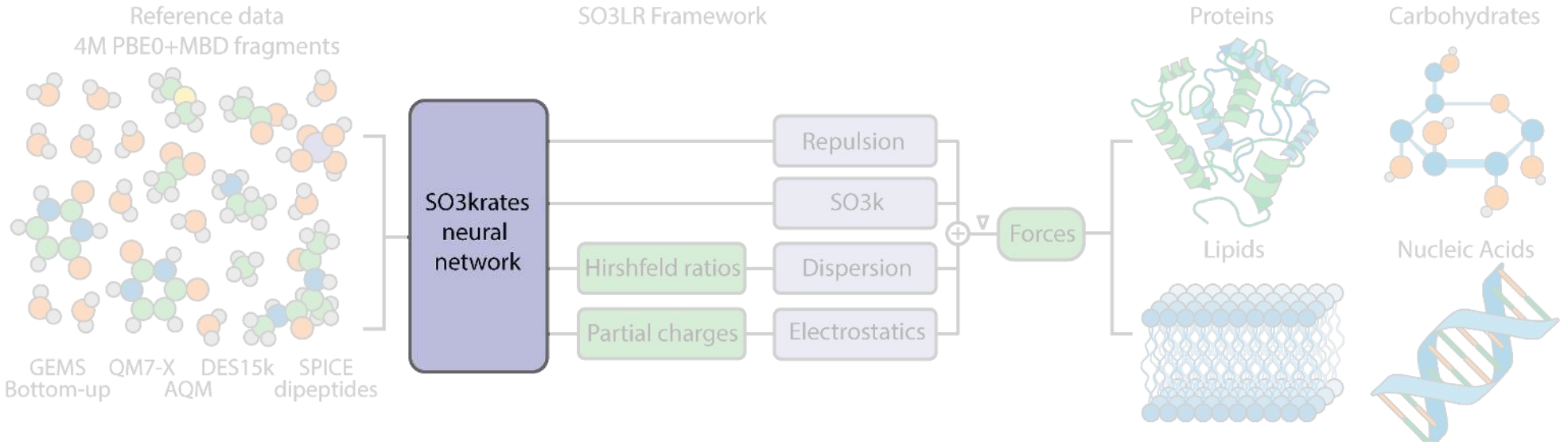
SO3krates – cornerstone of SO3LR

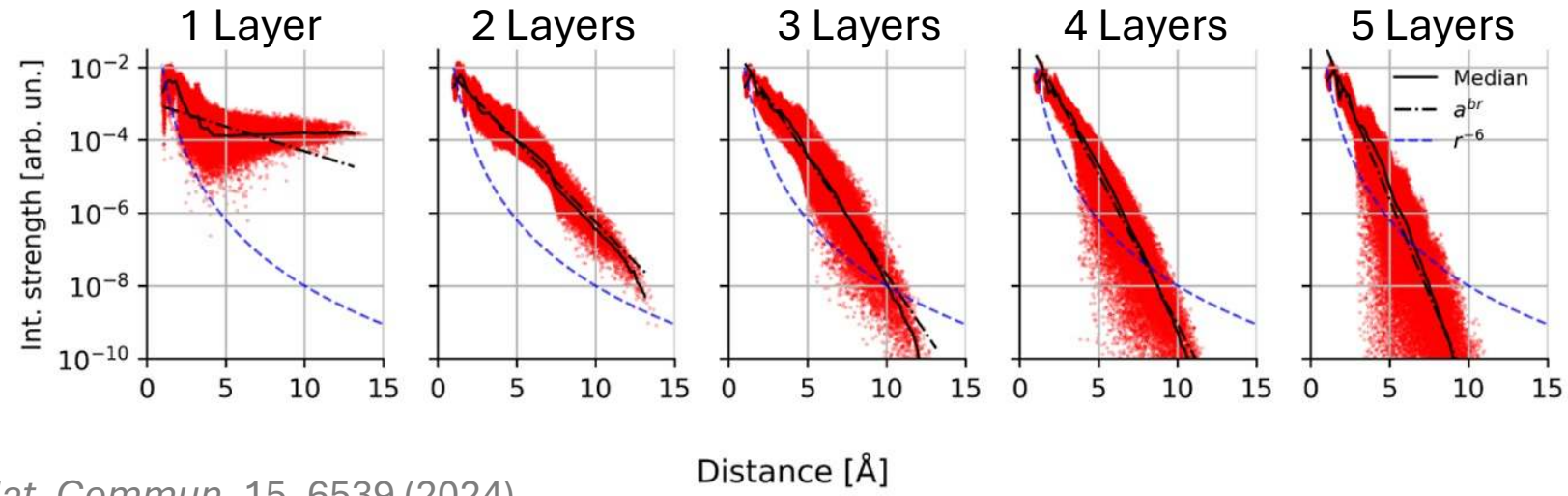




Frank et al., *Nat. Commun.* 15, 6539 (2024)
Esders, Schnake, Lederer, Kabylda et al., *JCTC* 21, 714 (2025)

SO3krates – cornerstone of SO3LR

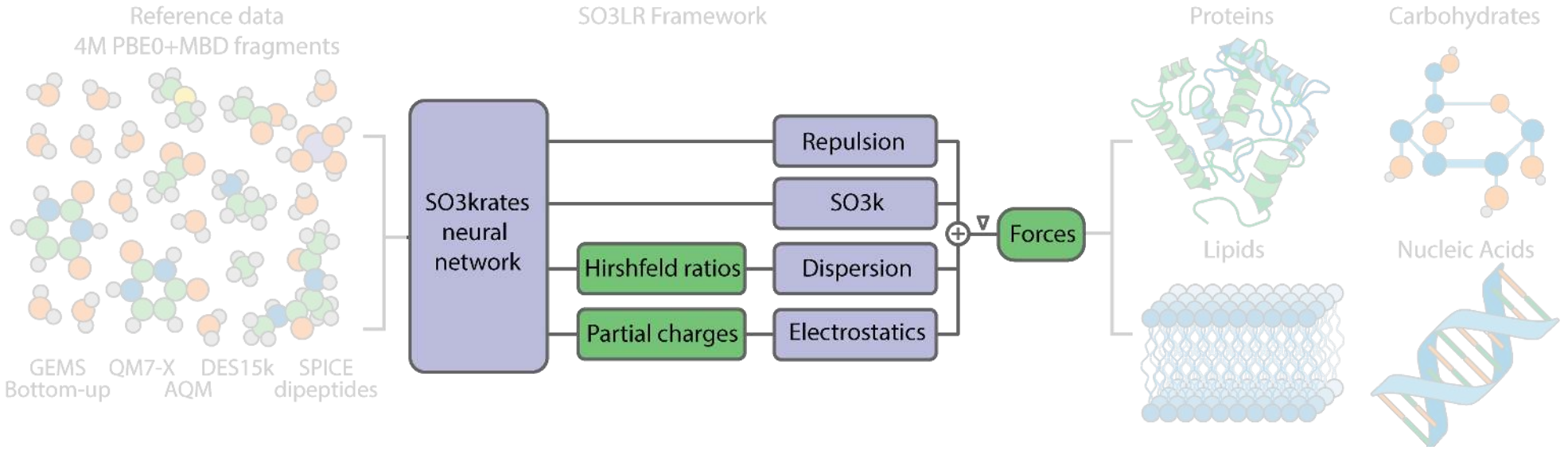




- MPNNs struggle to describe long-range interactions

Frank et al., *Nat. Commun.* 15, 6539 (2024)
Esders, Schnake, Lederer, Kabylda et al., *JCTC* 21, 714 (2025)

Universal Pairwise Force Fields



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

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

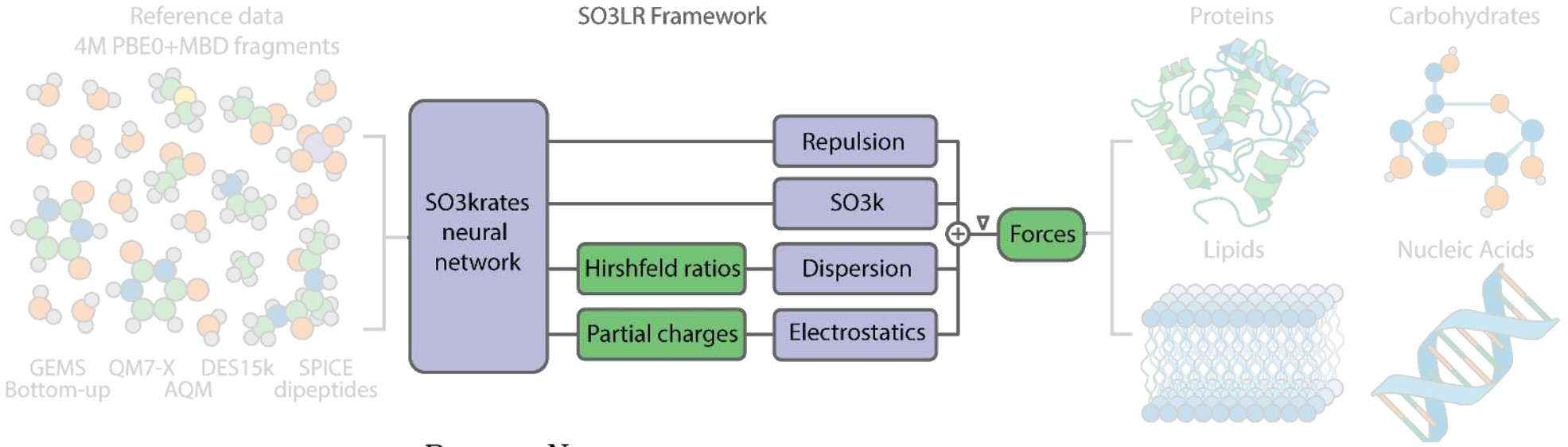
$$E_i, q_i, h_i = \text{SO3krates}\left(\mathcal{R}, \mathcal{Z}, Q_{\text{tot}}, S_{\text{tot}}\right)$$

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

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

Khabibrakhmanov et al., *JCTC* 19, 21, 7895 (2023)
Unke et al., *Nat. Commun.* 12, 7273 (2021)

SO3LR Training



Loss function:

$$\mathcal{L} = \frac{\lambda_F}{B}\sum_{b=1}^{B}\frac{1}{N_b}\sum_{i=1}^{N_b}\left|\left|\vec{F}_{i,\text{true}} - \vec{F}_{i,\text{pred}}\right|\right|_2^2$$
$$+ \frac{\lambda_\mu}{B}\sum_{b=1}^{B}\left|\left|\vec{\mu}_{b,\text{true}} - \vec{\mu}_{b,\text{pred}}\right|\right|_2^2$$
$$+ \frac{\lambda_h}{B}\sum_{b=1}^{B}\frac{1}{N_b}\sum_{i=1}^{N_b}\left(h_{i,\text{true}} - h_{i,\text{pred}}\right)^2,$$

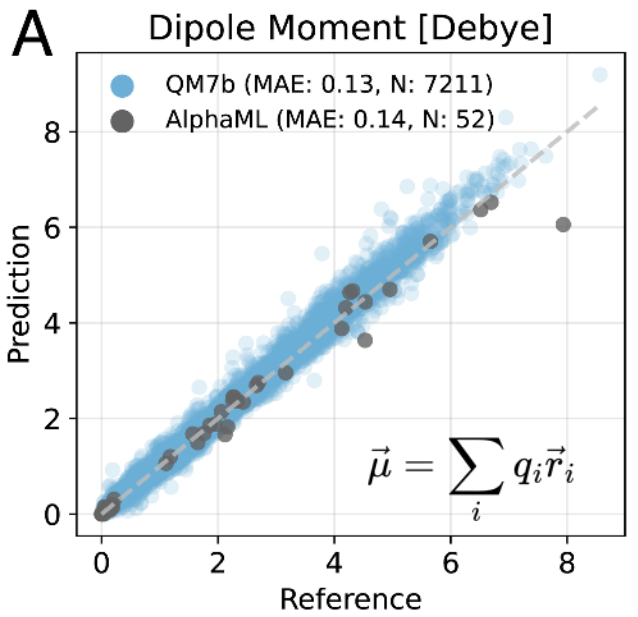
Kabylda et al., *JACS, 147, 37, 33723* (2025)

Static benchmarks

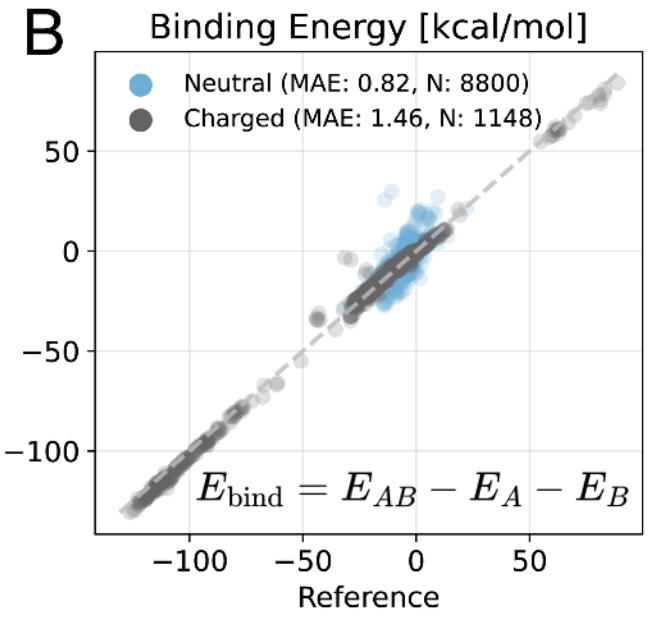
TABLE I. Root Mean Square Error of the model on various test sets. Force (eV/Å), dipole moment vector (e × Å), and Hirshfeld ratios. Dash indicates no data.

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	-
$AcAla_3NMe$	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

QM7b and AlphaML
LR-CCSD/d-aug-cc-pVDZ



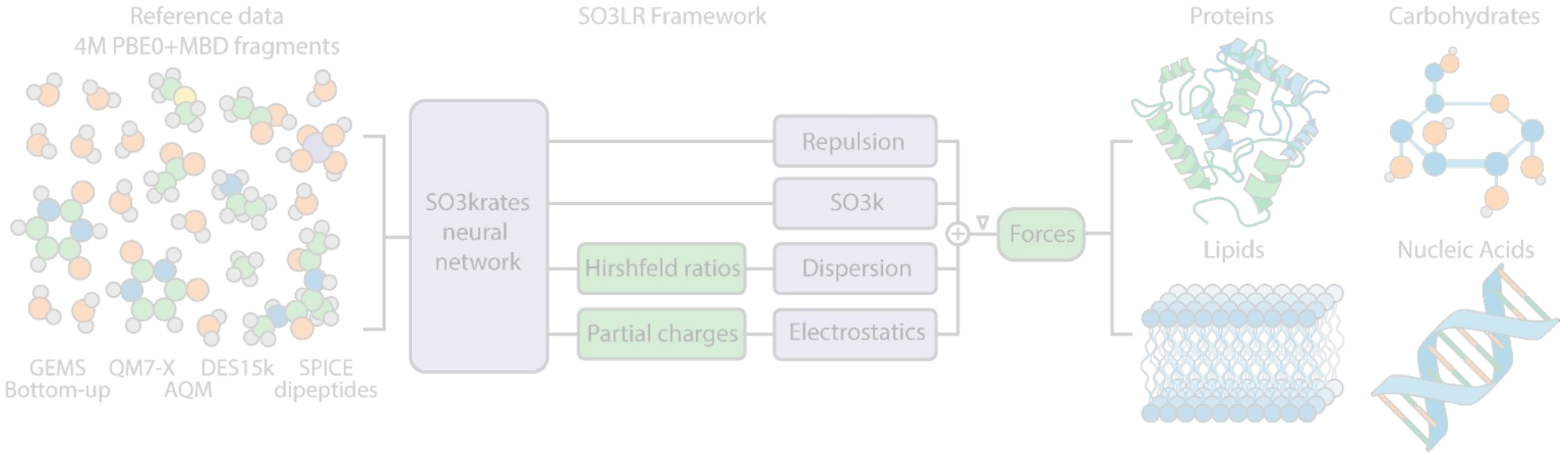
SAPT10k
SAPT2+(3)(CCD)/aug-cc-pVTZ

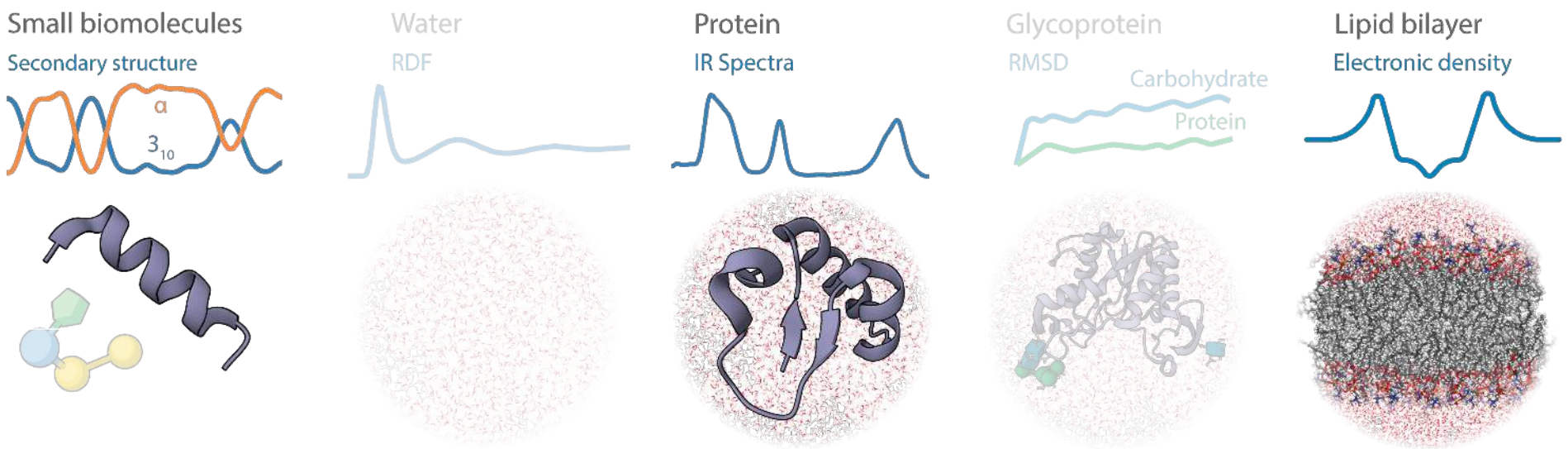


- Accurate dipole moment and binding energy predictions
- Rare outliers include dimers like ClF, $P(CNO)_3$, and PH_2NO_2 complexes

Yang et al., *Sci. Data* 6, 152 (2019)
Villot and Lao, *J. Chem. Phys.* 160, 184103 (2024)

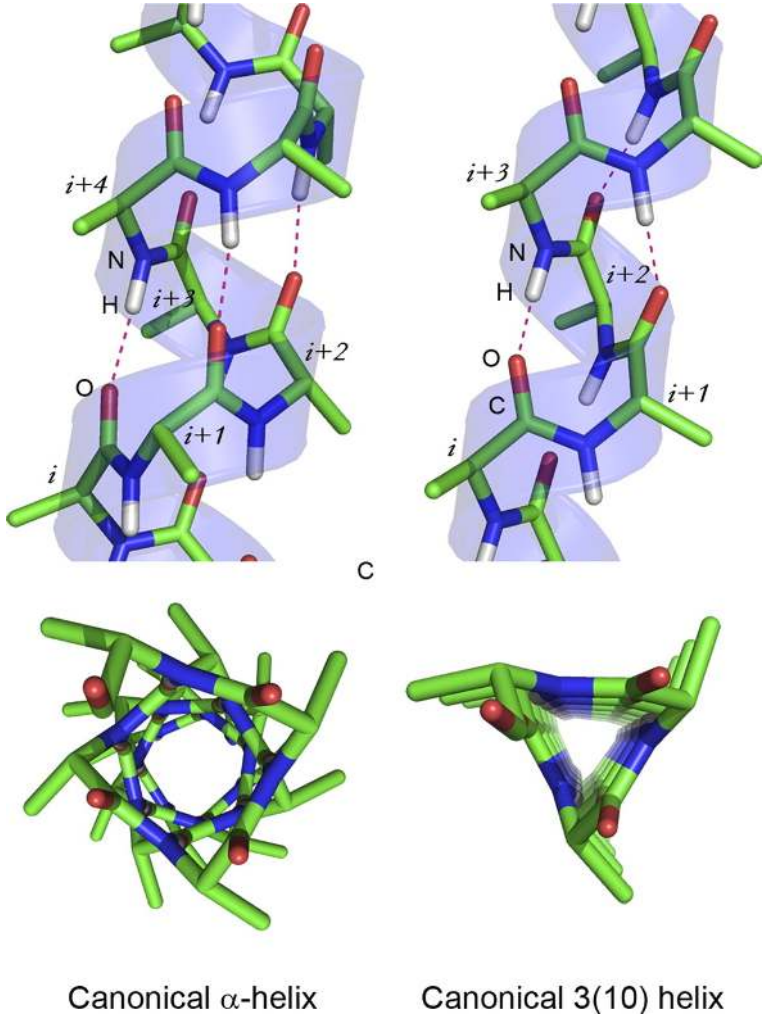
(Bio)Molecular Simulations with SO3LR





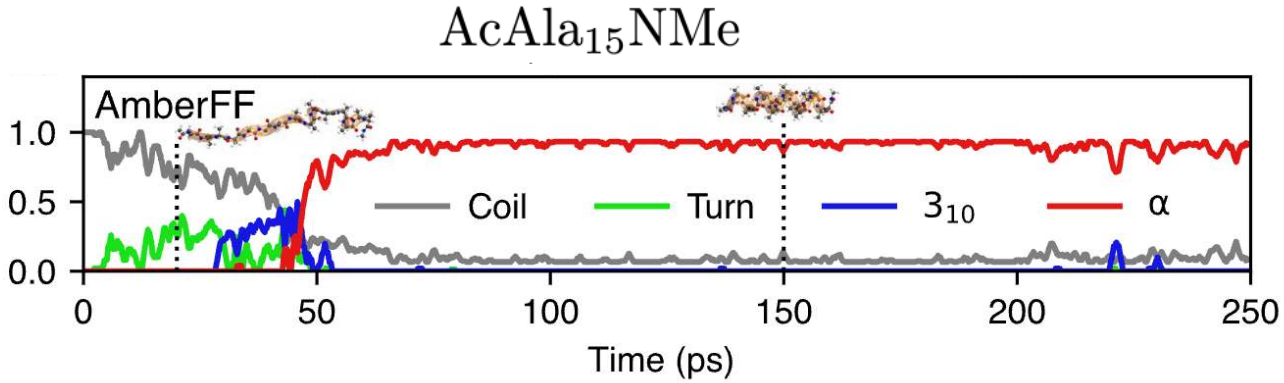
Kabylda et al., *JACS, 147, 37, 33723* (2025)

Folding of polyalanine



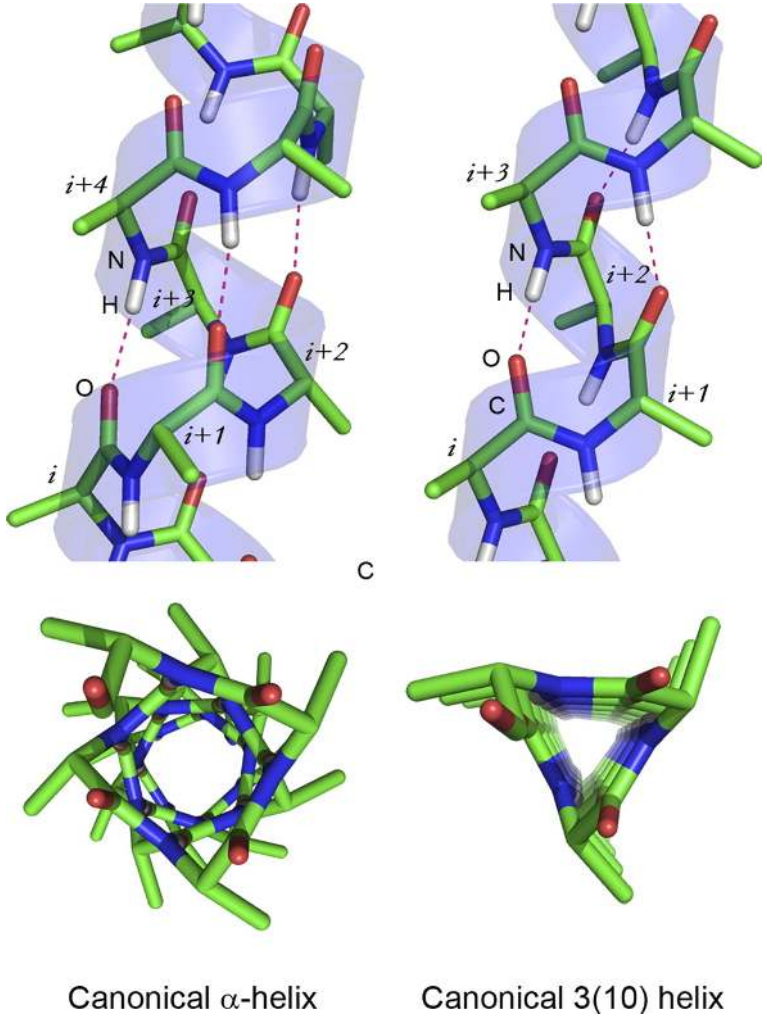
Canonical α-helix Canonical 3(10) helix

Vieira-Pires et al., *J. Gen. Physiol.* 136, 585 (2010)
Bolin and Millhauser, *Acc. Chem. Res.* 32, 12, 1027 (1999)

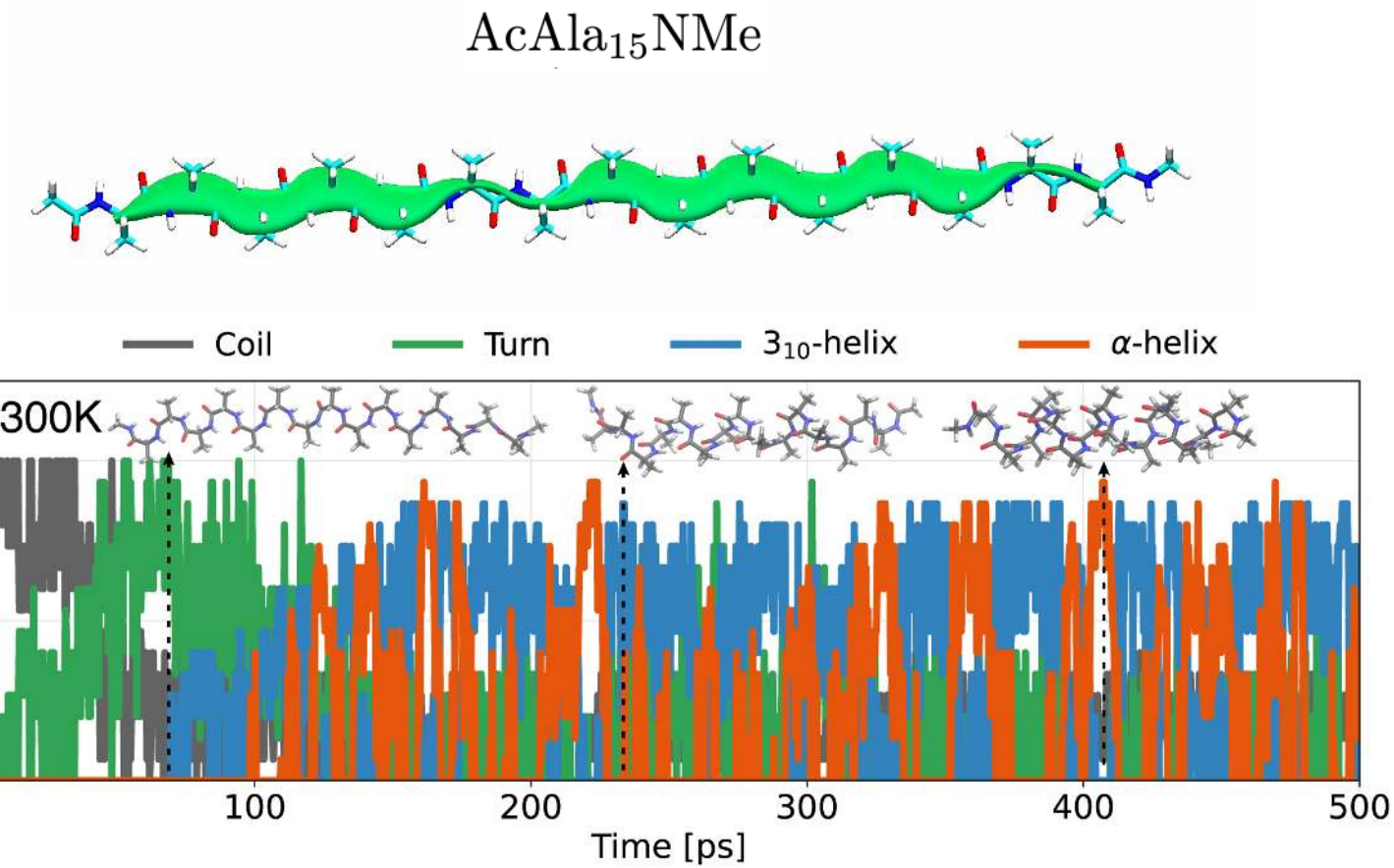


Unke et al., *Sci. Adv.* eadn4397 (2024)

Folding of polyalanine



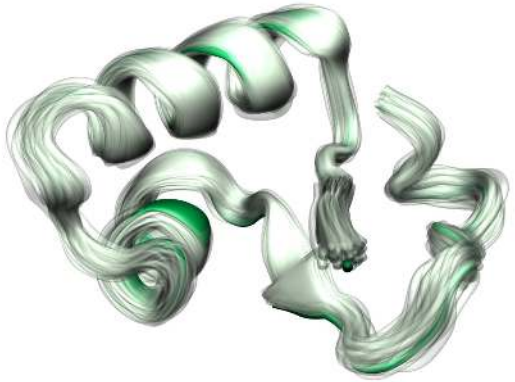
Canonical α-helix Canonical 3(10) helix



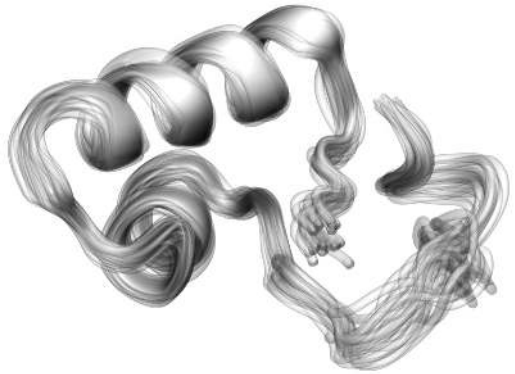
- Folding and alterations between alpha and 3_{10} forms
- Delicate interplay of hydrogen bonding, polarization, and dispersion interactions
- No system-specific Ala_{15} data

Vieira-Pires et al., *J. Gen. Physiol.* 136, 585 (2010)
Bolin and Millhauser, *Acc. Chem. Res.* 32, 12, 1027 (1999)

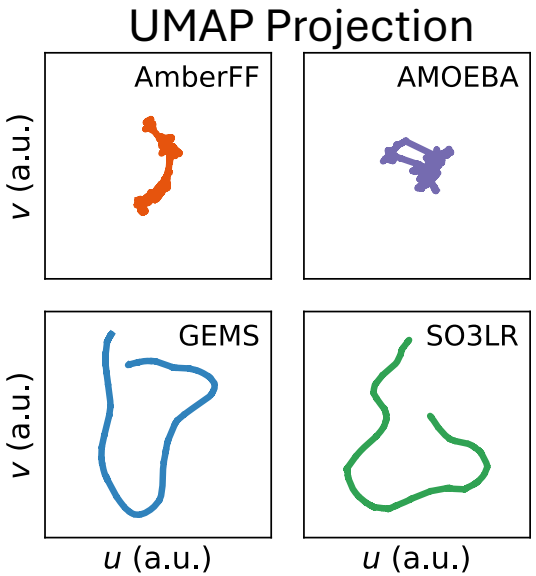
Crambin

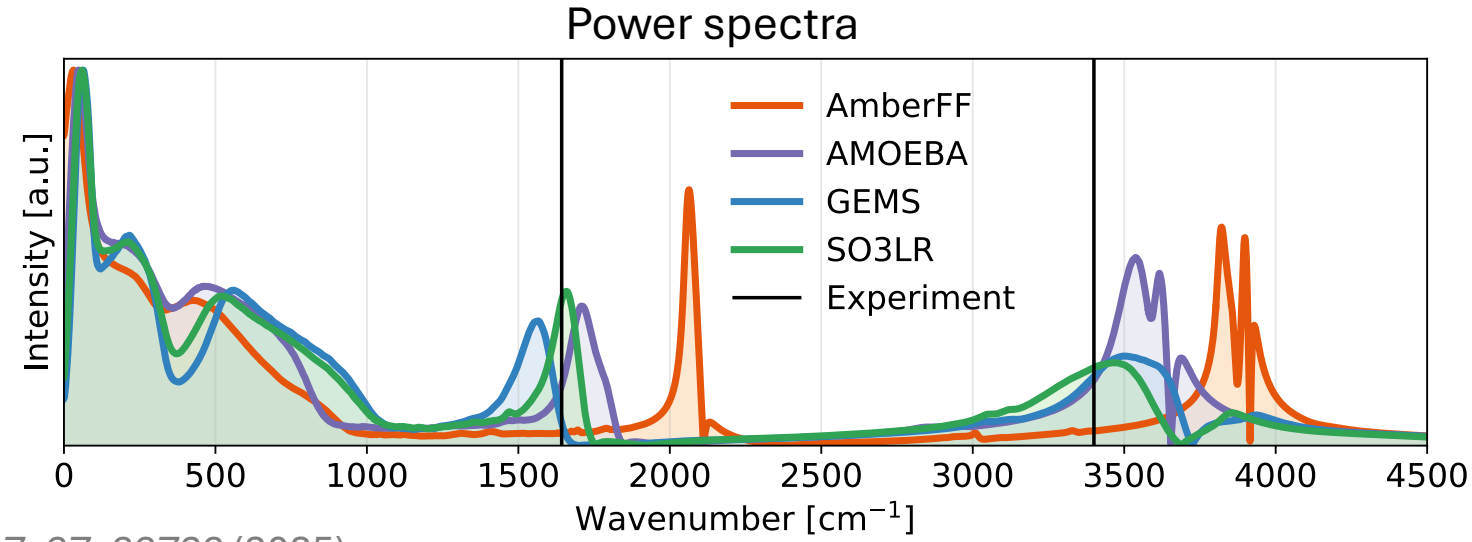


SO3LR MD Trajectory
25k atoms, 3 ns



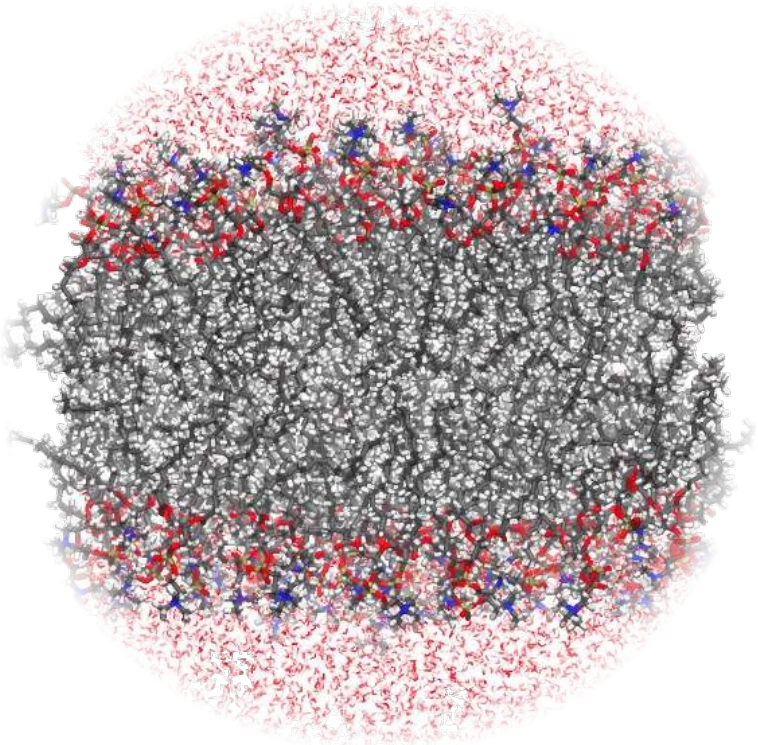
Experimental
structures





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

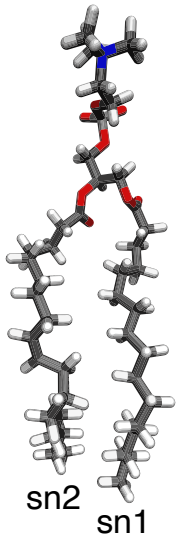
POPC Lipid bilayer

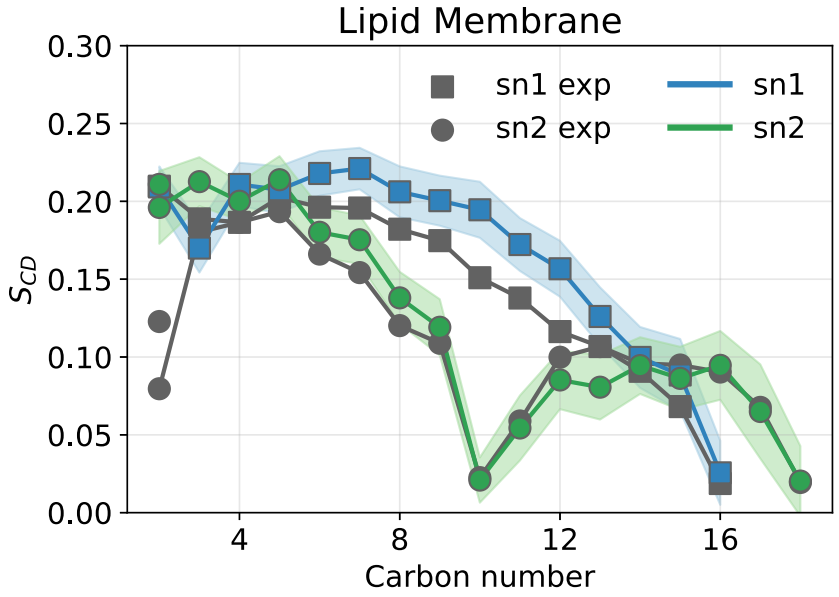


33k atoms
500 ps

TABLE II. POPC Lipid Bilayer Structural Properties: bilayer thicknesses D_{HH} (Å) and area per lipid (Å^2).

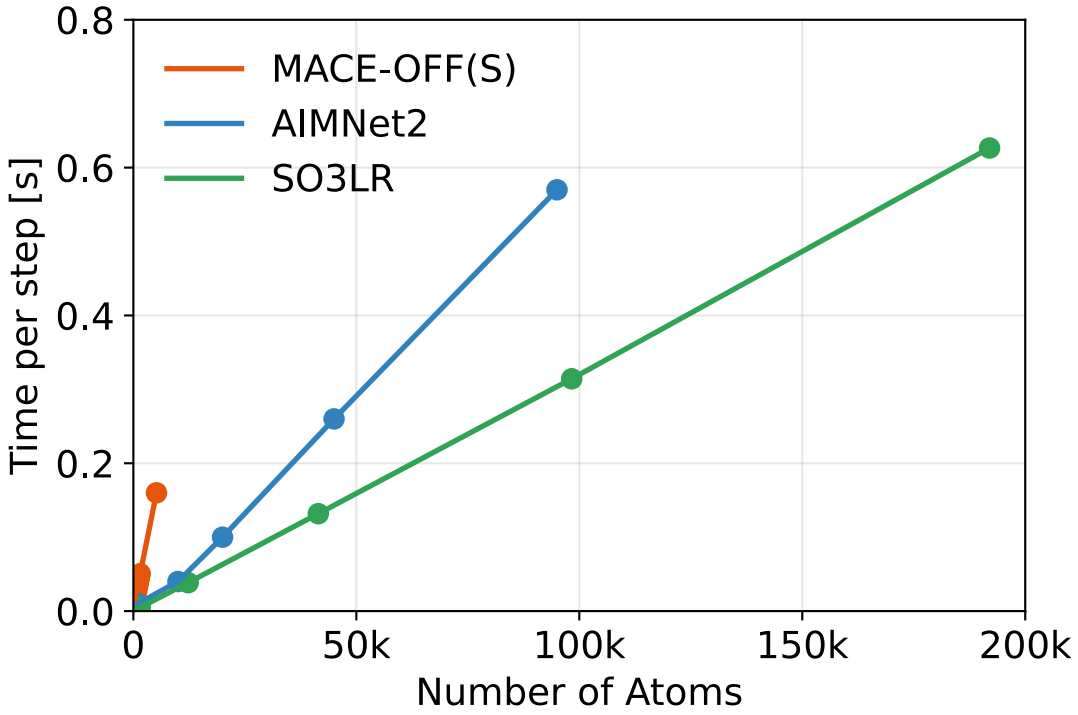
	D_{HH}	S/lipid
Experiment [76]	36.5	64.3±1.3
Lipid21 [77]	38.50±0.20	63.92±0.09
CHARMM36/LJ-PME [78]	37.3±0.30	65.4±0.5
SO3LR	37.0	58.0±0.1





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

Single GPU performance

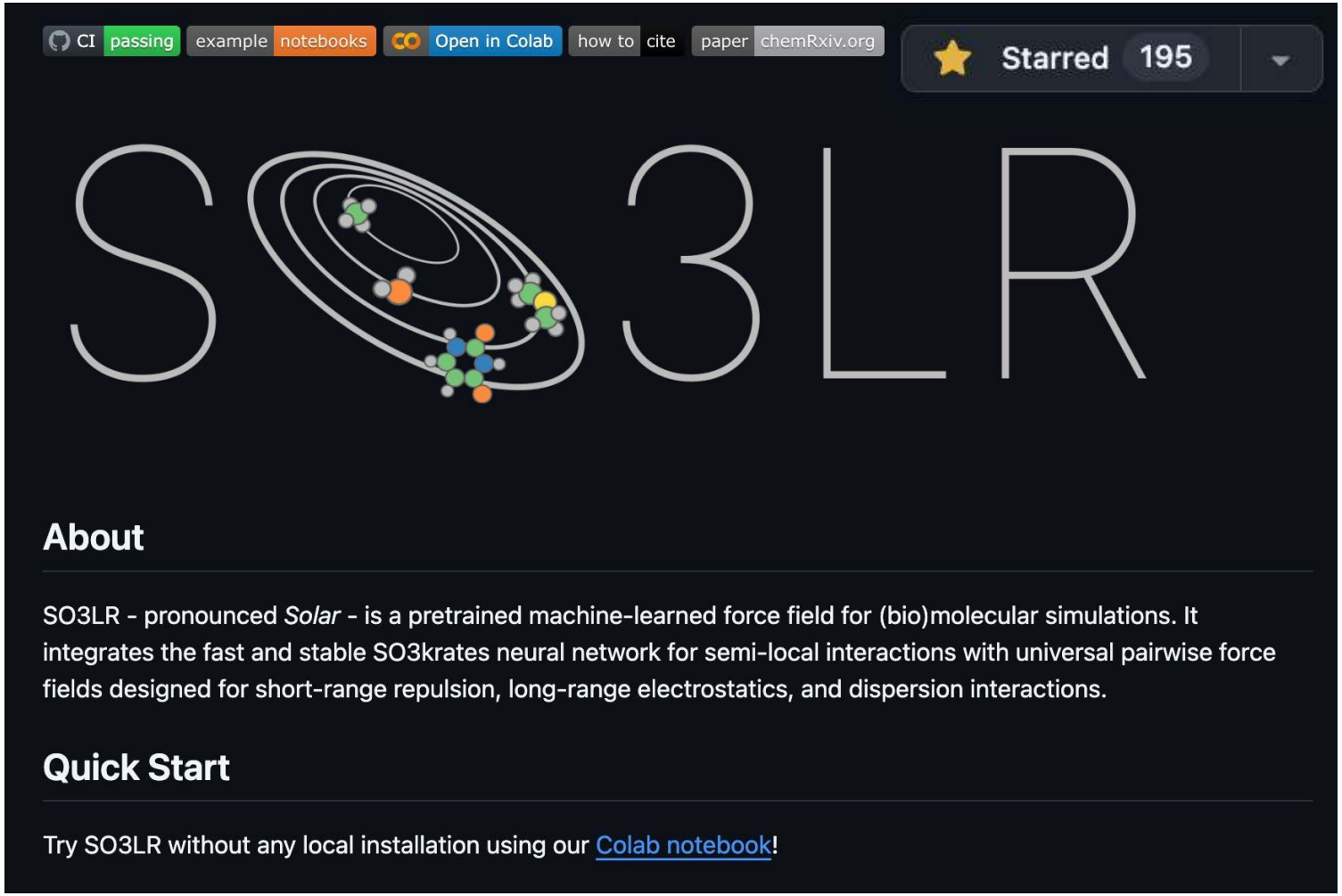


- ~3 μs/atom/step on a single H100 GPU (2.6 ns/day for a 10,000-atom system)
- approaching sizes and timescales relevant for realistic biomolecules, while preserving QM accuracy

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

Model, Data, and Code

github.com/general-molecular-simulations/so3lr



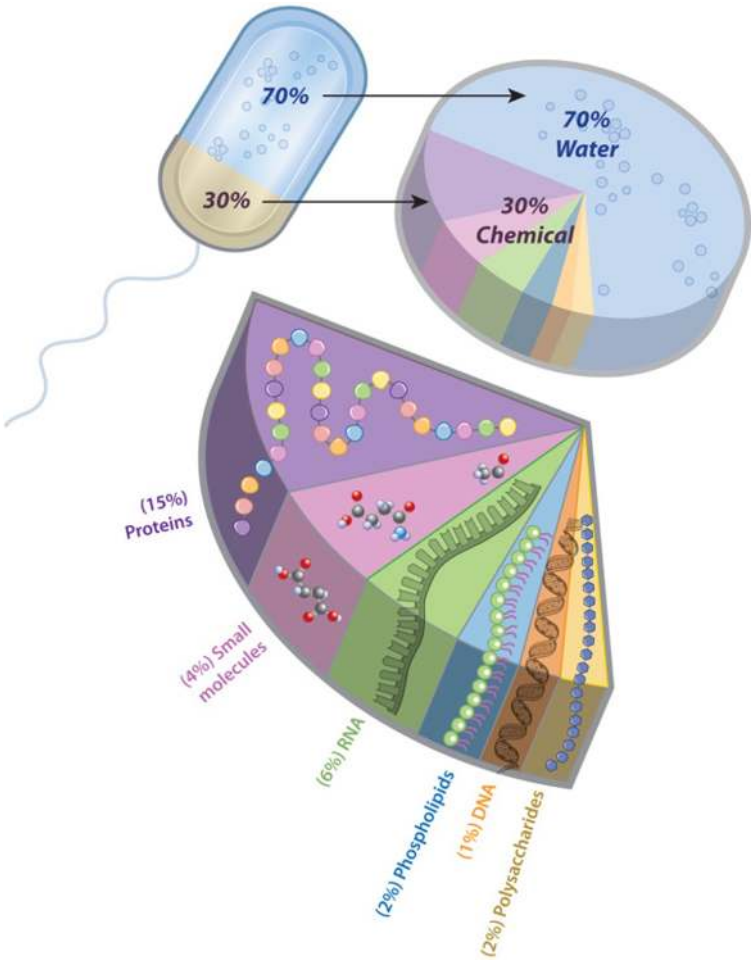
- ✓ Works with ASE/JAX-MD
- ✓ Google Colab/Tutorials
- ✓ Command Line Interface
  so3lr nvt --input mol.xyz

Outline

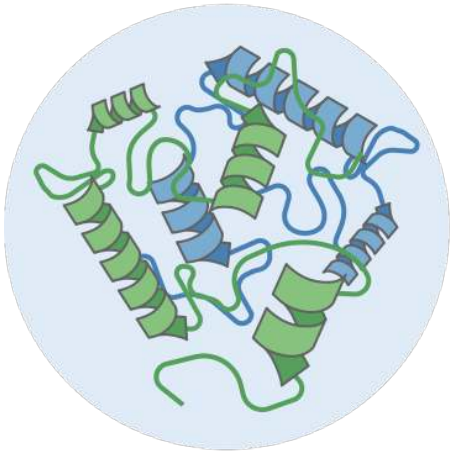
- *Introduction*
- *SO3LR: General-purpose MLFF for organic matter*
- ***QCell: QM dataset spanning diverse biomolecular fragments***
- *Conclusions & Outlook*

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)

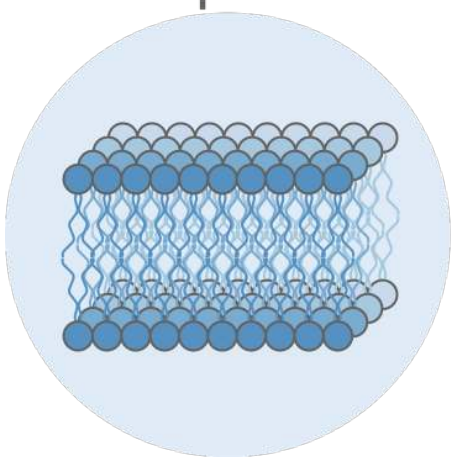
Major classes of biomolecules



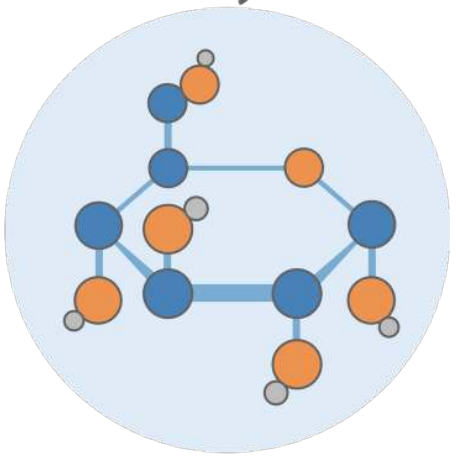
Proteins



Lipids



Carbohydrates



Nucleic Acids

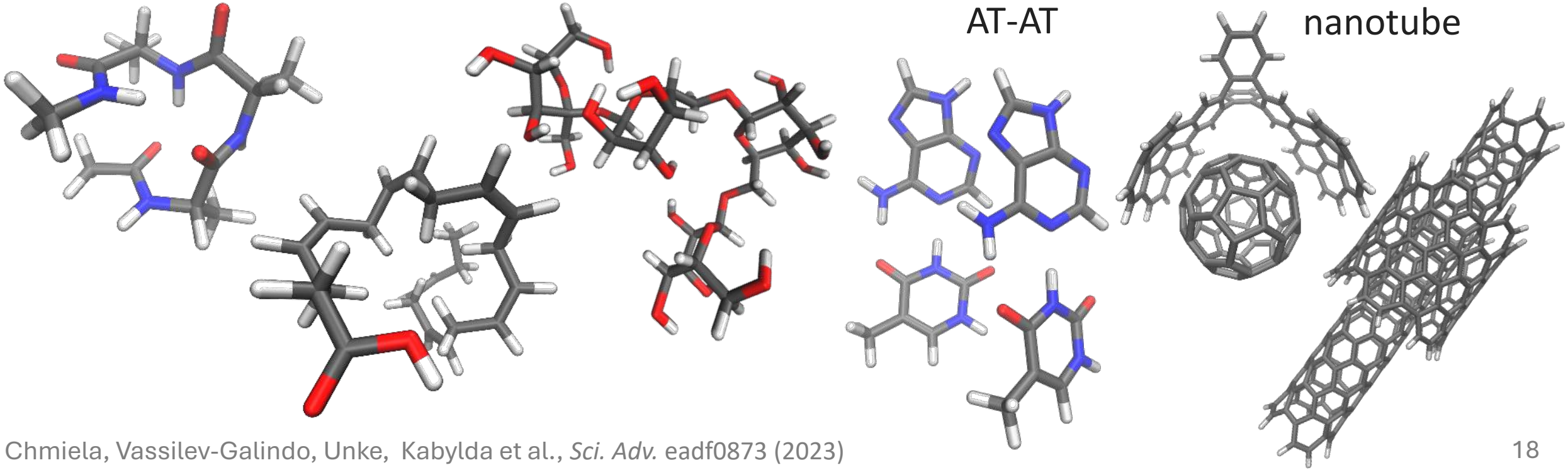
Alberts et al., Molecular Biology of the Cell

MD22 Dataset

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

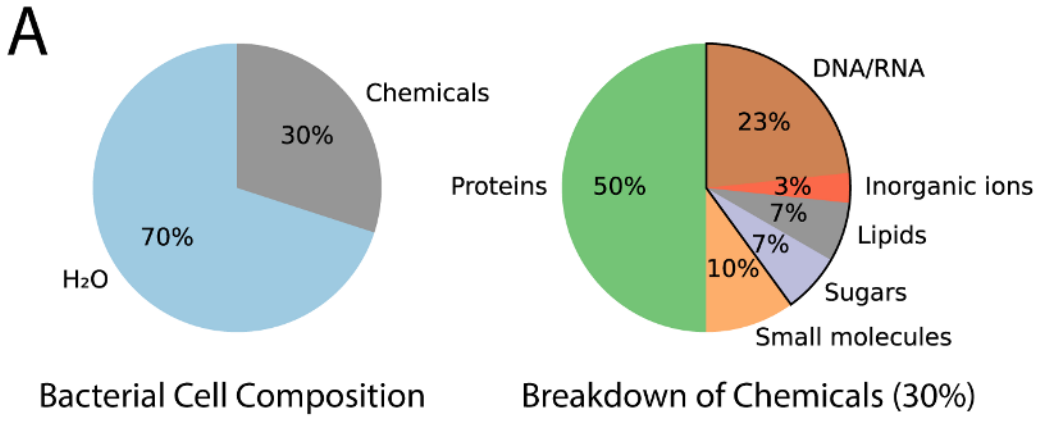
FHI-aims
The ab initio materials simulation package

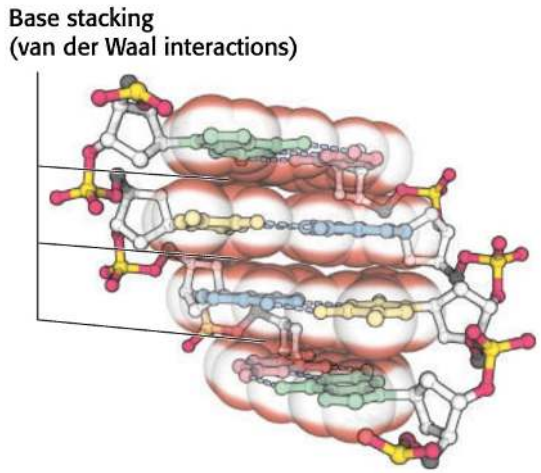
Proteins	Lipids	Carbohydrates	Nucleic acids	Supramolecules
Tetrapeptide	Fatty acid	Tetrasaccharide	DNA base pairs	Buckyball Catcher
Ac-Ala3-NHMe	DHA	Stachyose	AT-AT-CG-CG	Double-walled
			AT-AT	nanotube

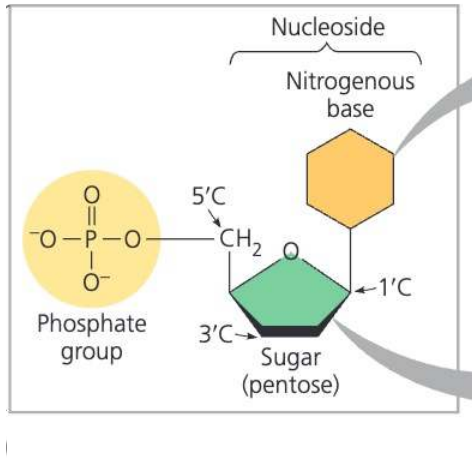


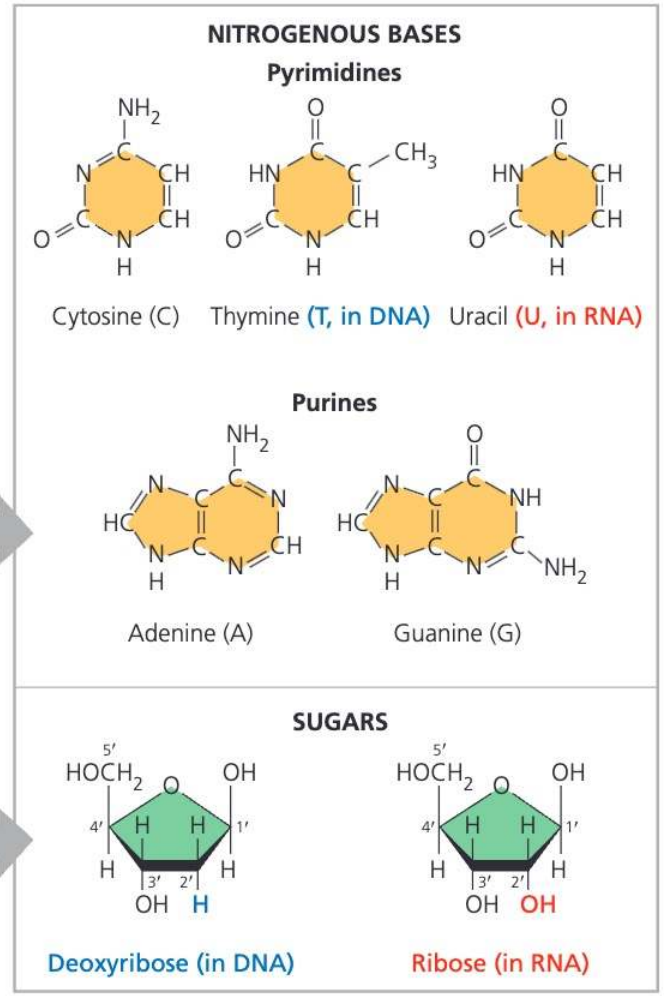
Chmiela, Vassilev-Galindo, Unke, Kabylda et al., *Sci. Adv.* eadf0873 (2023)

QCell: Biomolecular Building Blocks





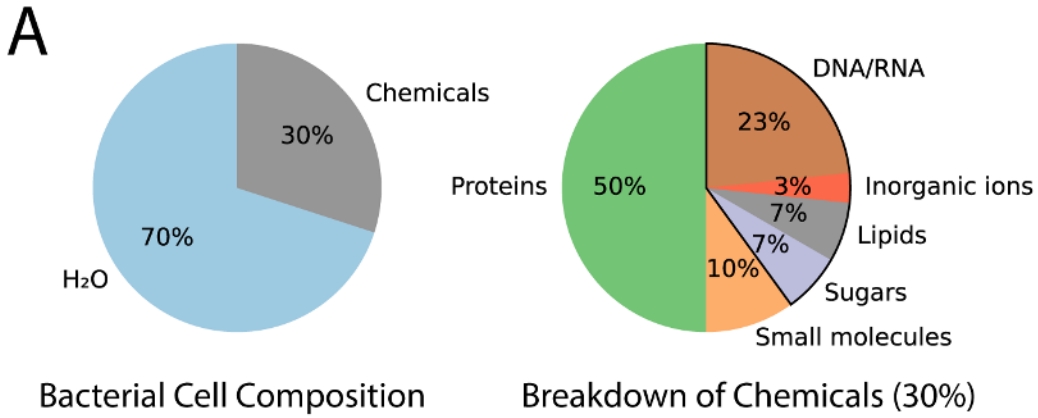


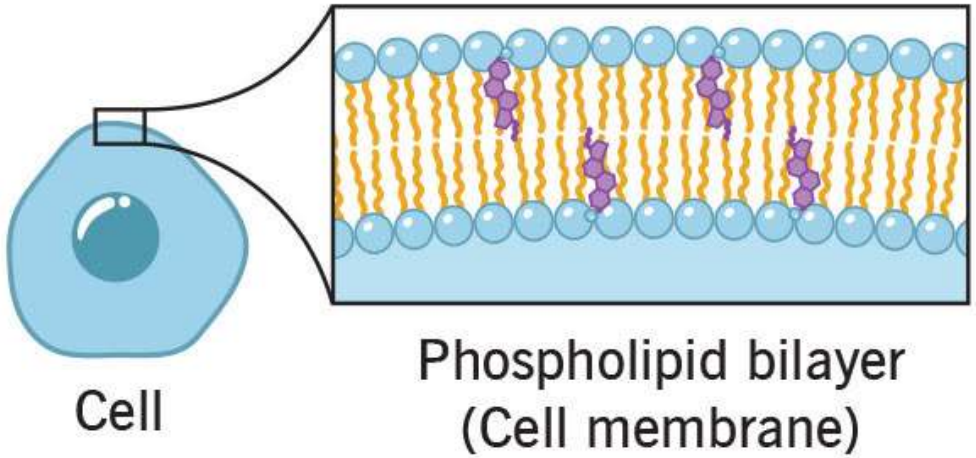


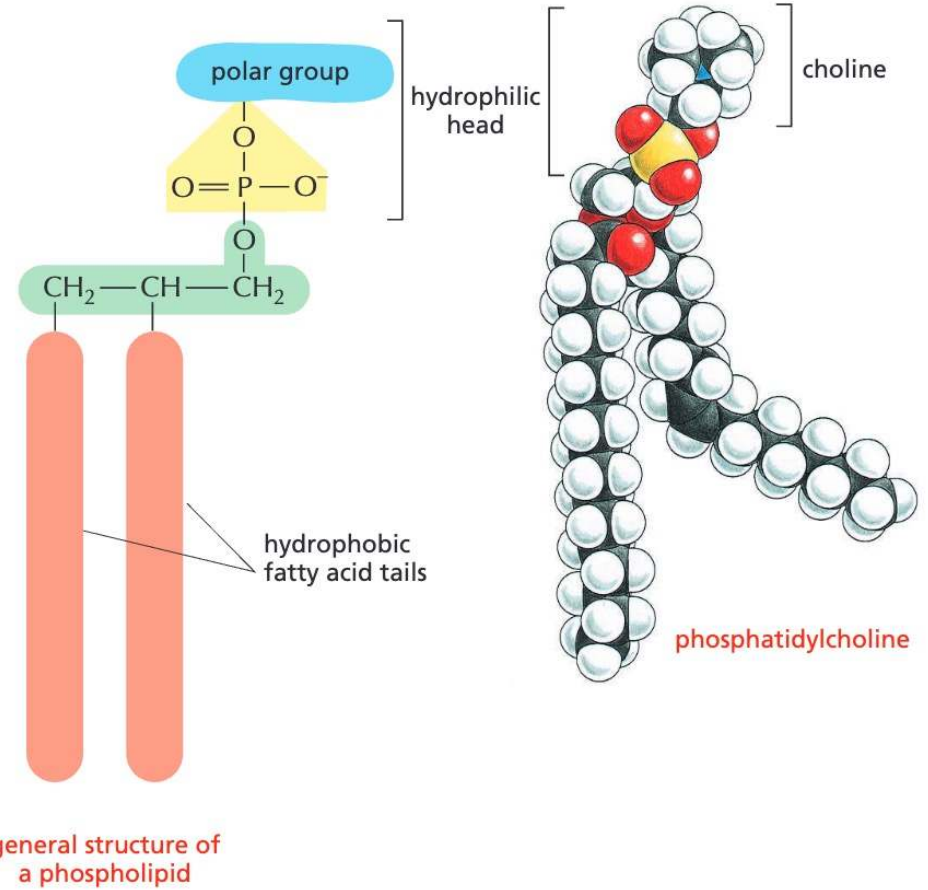
- Gas-phase RNA fragments
- Solvated double-helical DNA dimers/trimers in canonical A, B, and Z forms

Alberts et al., Molecular Biology of the Cell

QCell: Biomolecular Building Blocks



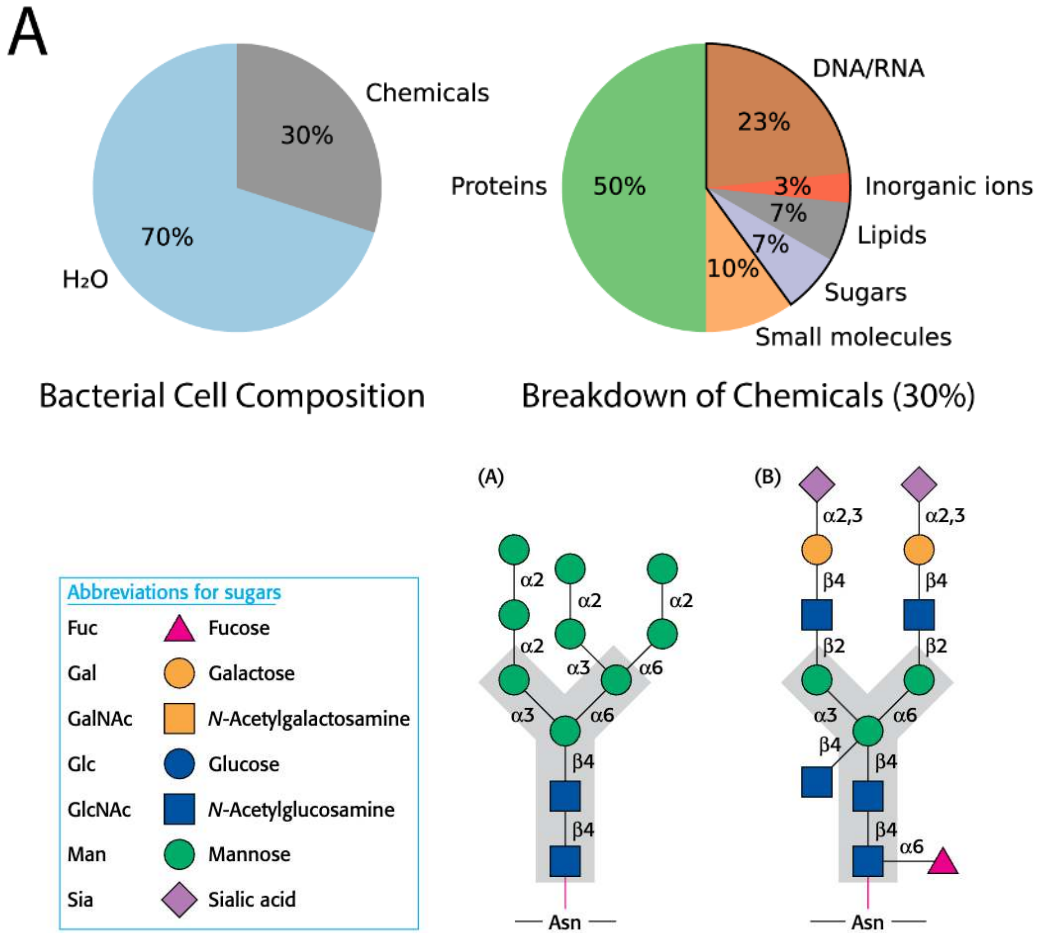


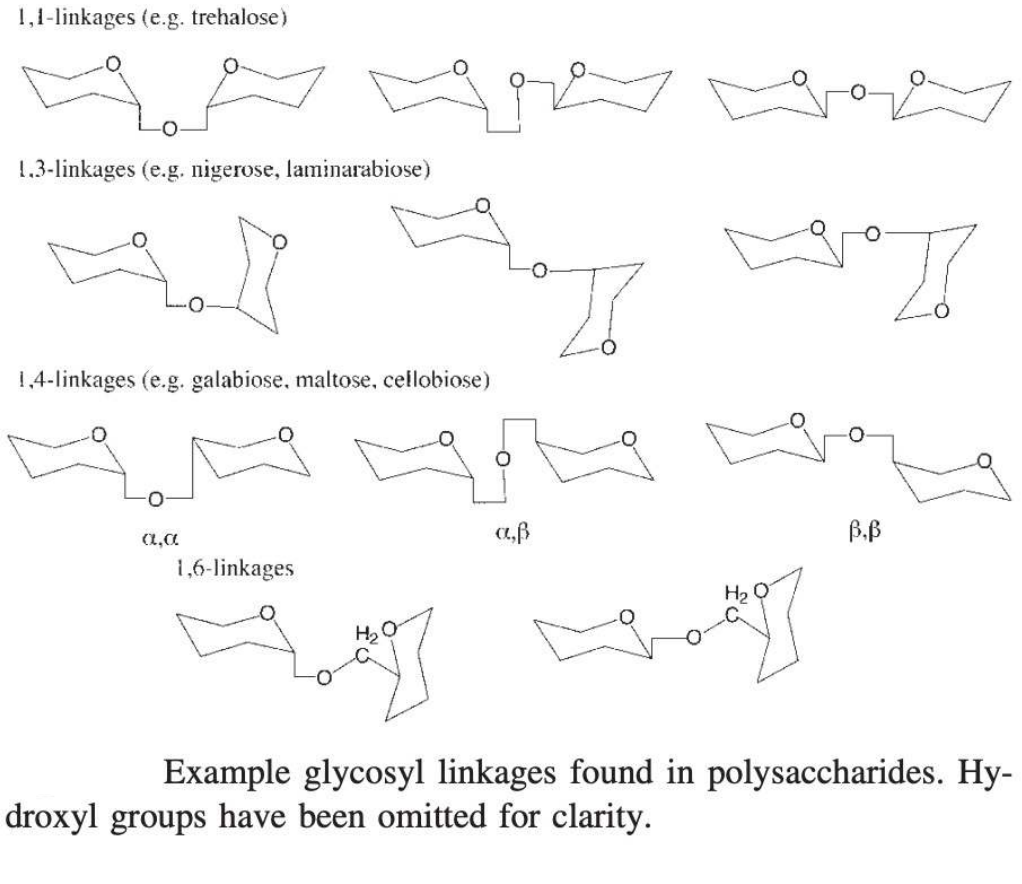


- POPC, POPE, POPG, and POPS 1-2-3-mers (with cholesterol)

Alberts et al., Molecular Biology of the Cell

QCell: Biomolecular Building Blocks



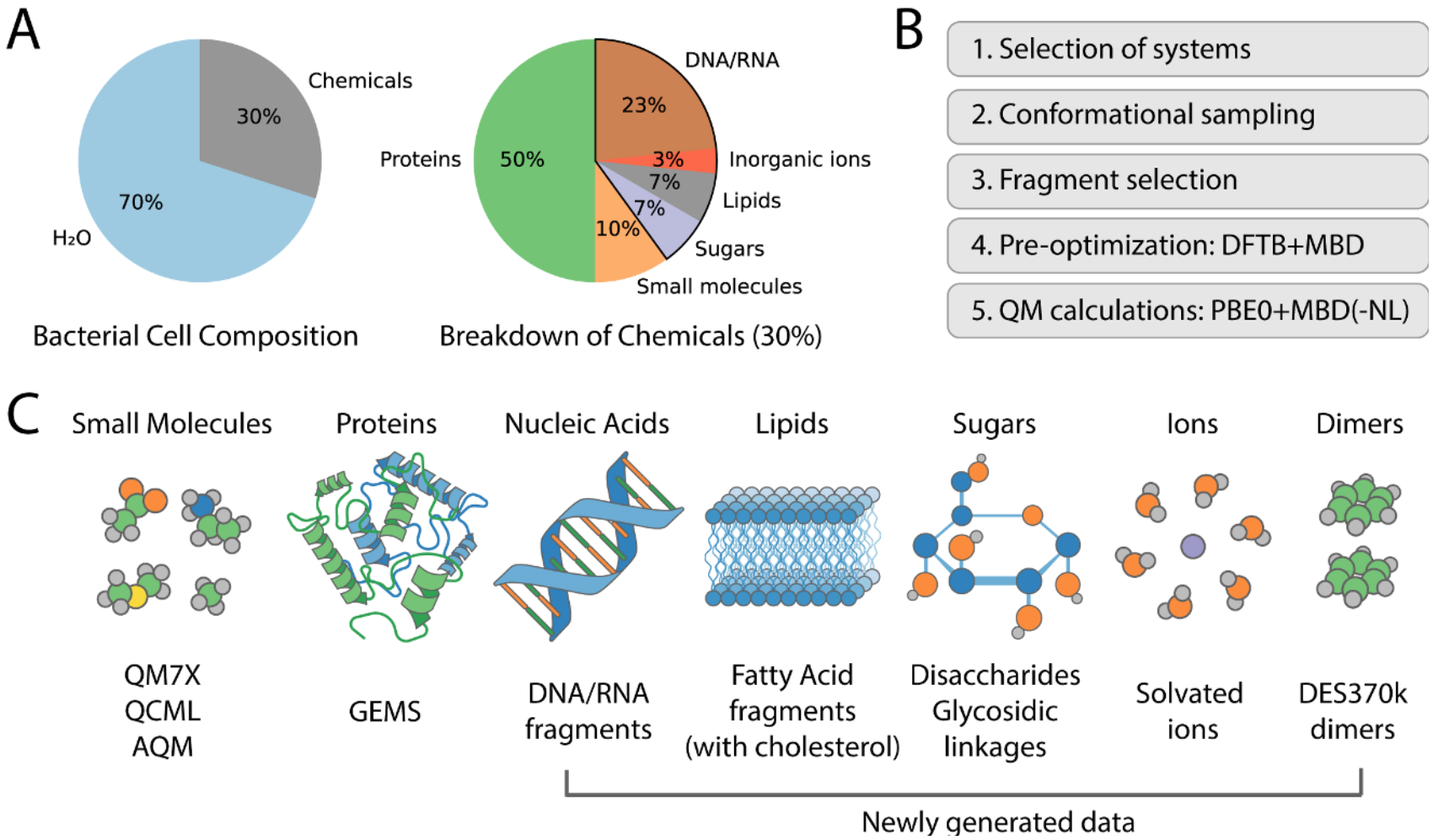


Example glycosyl linkages found in polysaccharides. Hydroxyl groups have been omitted for clarity.

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

Alberts et al., Molecular Biology of the Cell

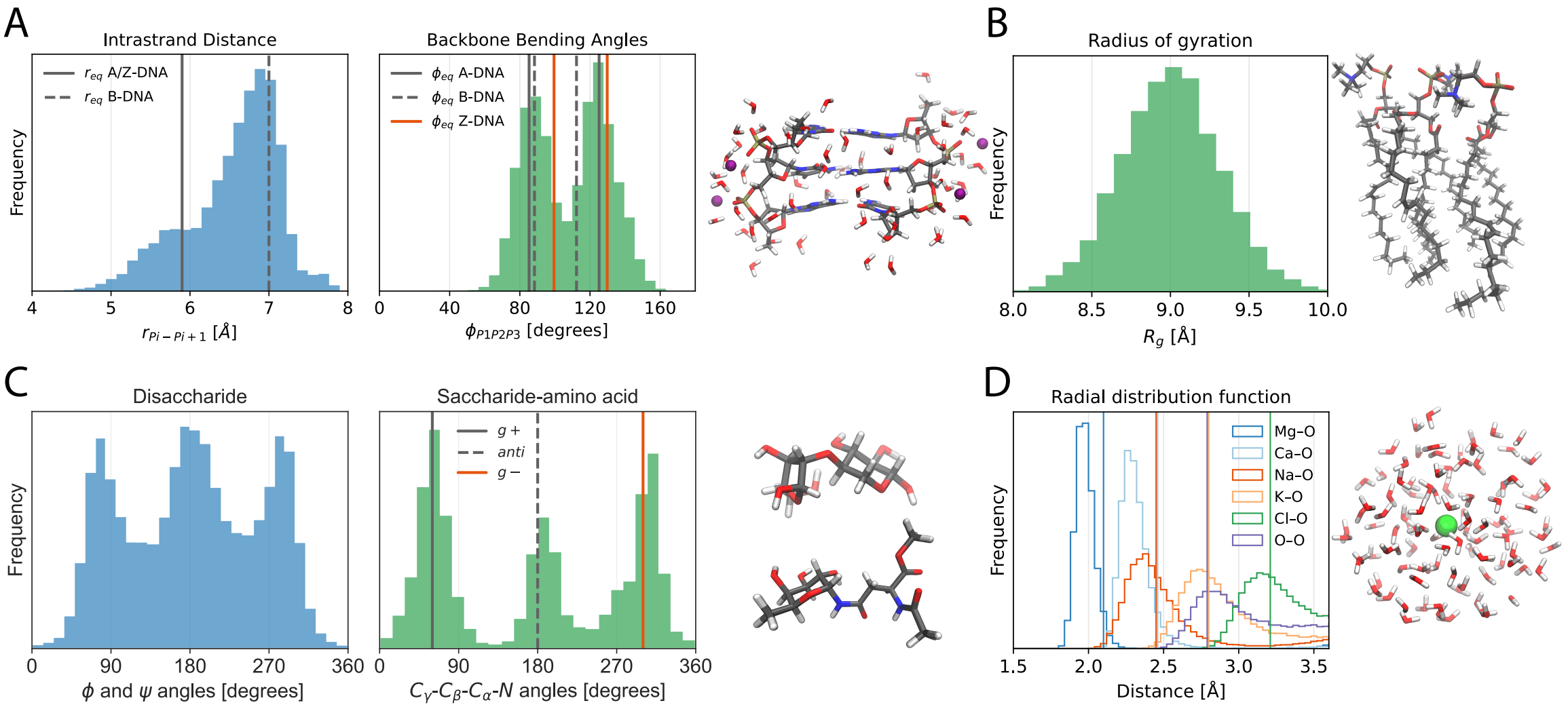
QCell: Biomolecular Building Blocks



- 0.5M -> Total: 41M (GEMS, QM7-X, QCML, AQM)

Kabylda et al., *AI Sci* 2, 025003 (2026)
Alberts et al., Molecular Biology of the Cell

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

UNIVERSITÉ DU LUXEMBOURG

TECHNISCHE UNIVERSITÄT BERLIN

Google DeepMind

Luxembourg National Research Fund

HPC

MELUXINA
HIGH PERFORMANCE COMPUTING IN LUXEMBOURG

Outline

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

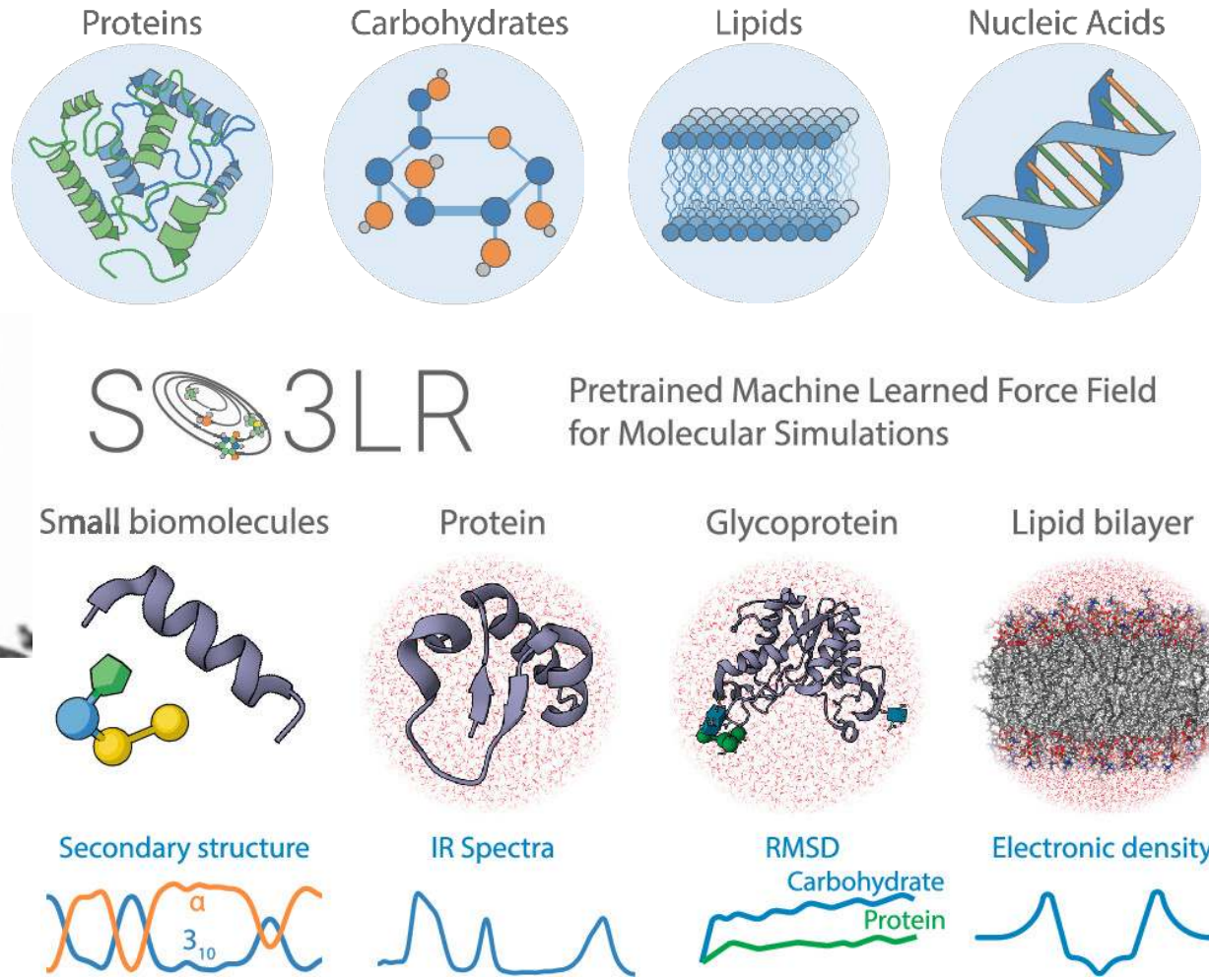
Conclusions & Outlook

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



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)