Friedrich-Alexander-Universität
Naturwissenschaftliche Fakultät

43384 – Digital Alchemy

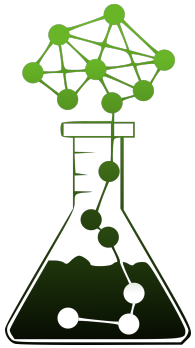
Unit 06 – Machine Learning in Chemistry

Prof. Dr. Carolin Müller

November 24, 2025

ML for Molecule Discovery

Overview

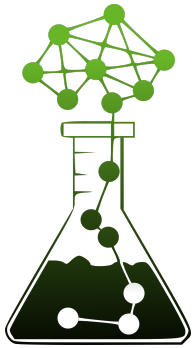


Why ML?

- Quantum chemistry is accurate but expensive
- Chemical space is astronomically large
- Experiments generate vast data (spectra, reactions, kinetics)

ML for Molecule Discovery

Overview

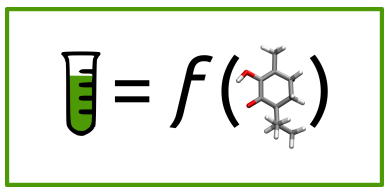


Why ML?

- Quantum chemistry is accurate but expensive
- Chemical space is astronomically large
- Experiments generate vast data (spectra, reactions, kinetics)

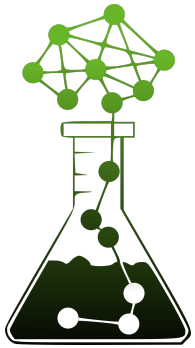
Core Applications

Molecular activity
& property prediction



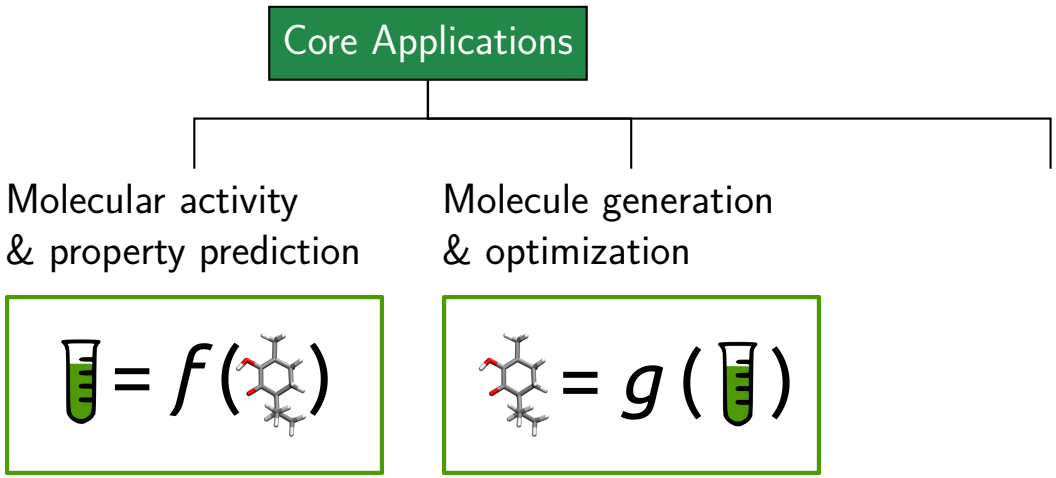
ML for Molecule Discovery

Overview



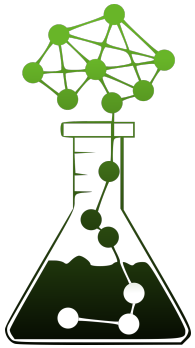
Why ML?

- Quantum chemistry is accurate but expensive
- Chemical space is astronomically large
- Experiments generate vast data (spectra, reactions, kinetics)



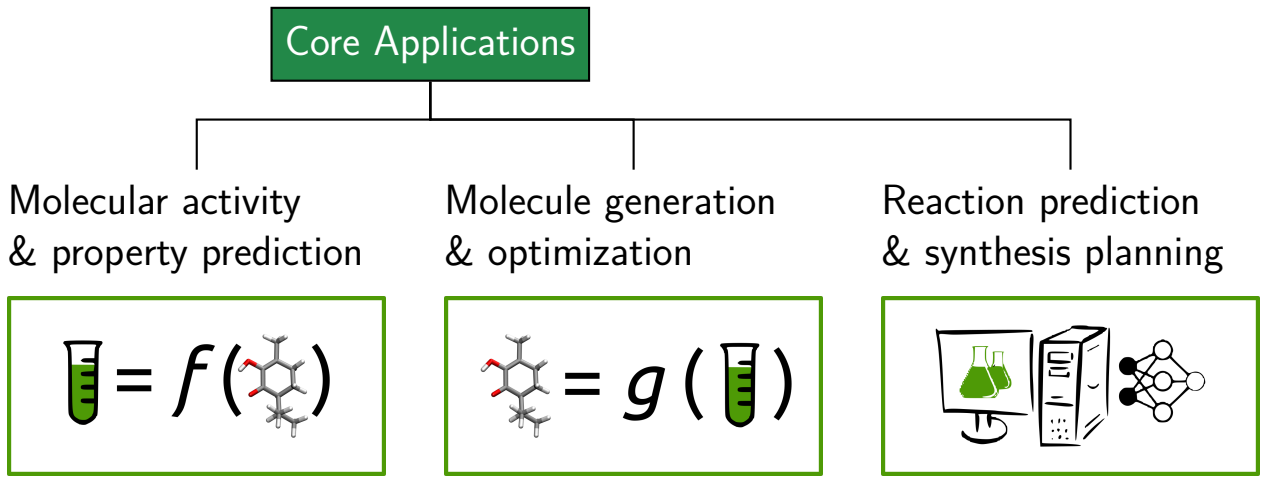
ML for Molecule Discovery

Overview



Why ML?

- Quantum chemistry is accurate but expensive
- Chemical space is astronomically large
- Experiments generate vast data (spectra, reactions, kinetics)



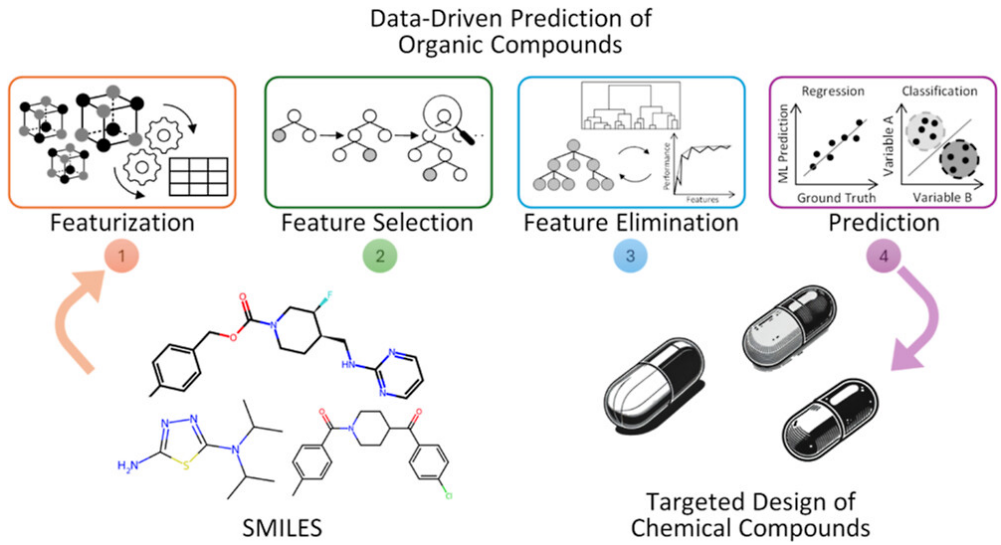
1. Molecular property prediction
2. Molecule generation and optimization
3. Reaction prediction and synthesis planning
4. Datasets for Molecule Discovery

1) Molecular Property Prediction

Motivation

Goal: Predict molecular properties directly from structure; map structure to property: $\mathbf{x} \rightarrow y$

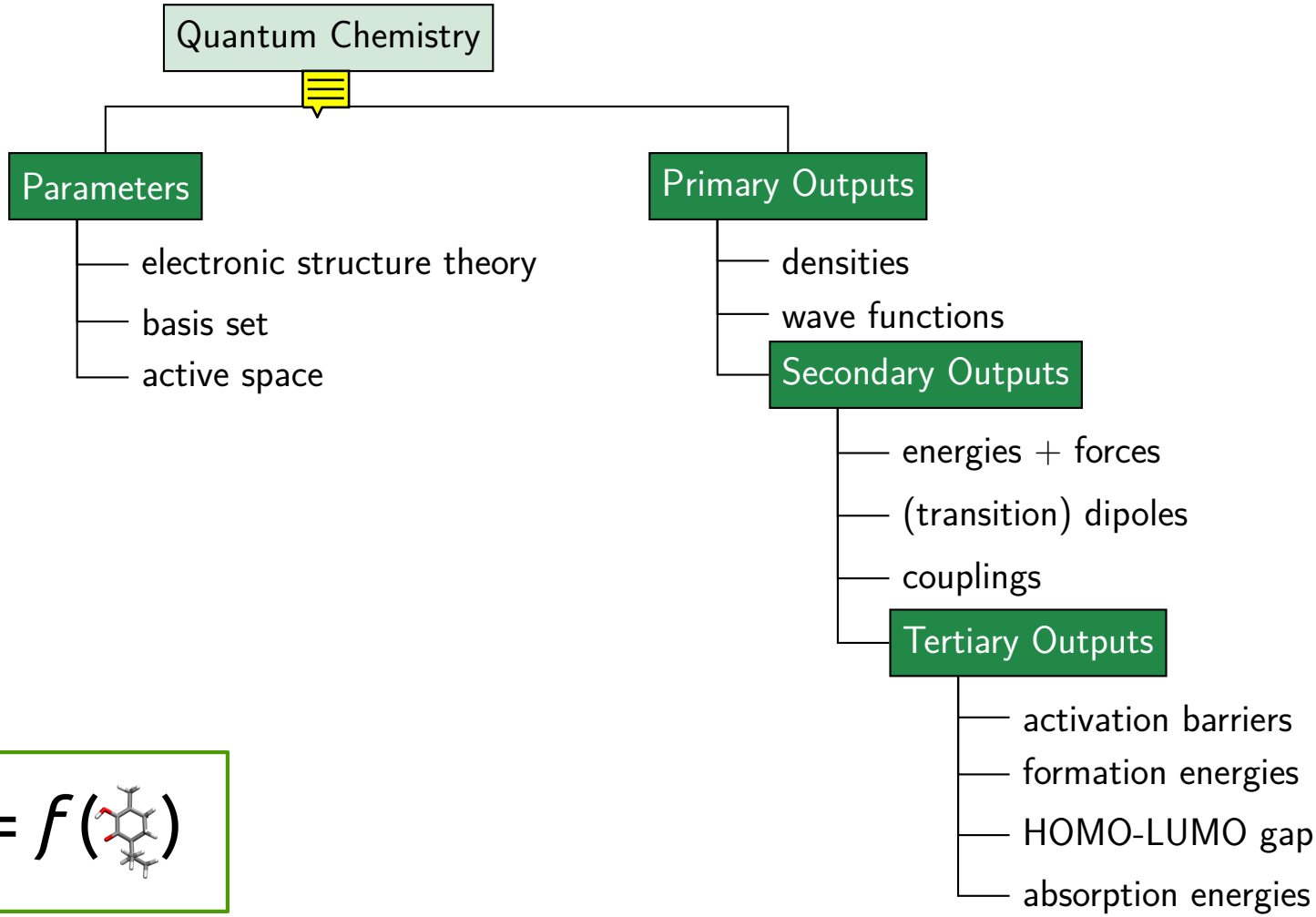
- **Screening & discovery:** Identify promising molecules/materials before synthesis (toxicity, activity, color, stability).
- **Design objectives:** Tune structures to achieve desired responses (*e.g.*, optical absorption, catalytic activity, redox stability).
- **Structure–function insight:** Understand how structural motifs influence measurable properties.
- **Accelerated computation:** Replace expensive *ab initio* property evaluations (energies, spectra, reaction barriers).



J. Cole *et al.*, 10.1021/acs.jcim.4c01862

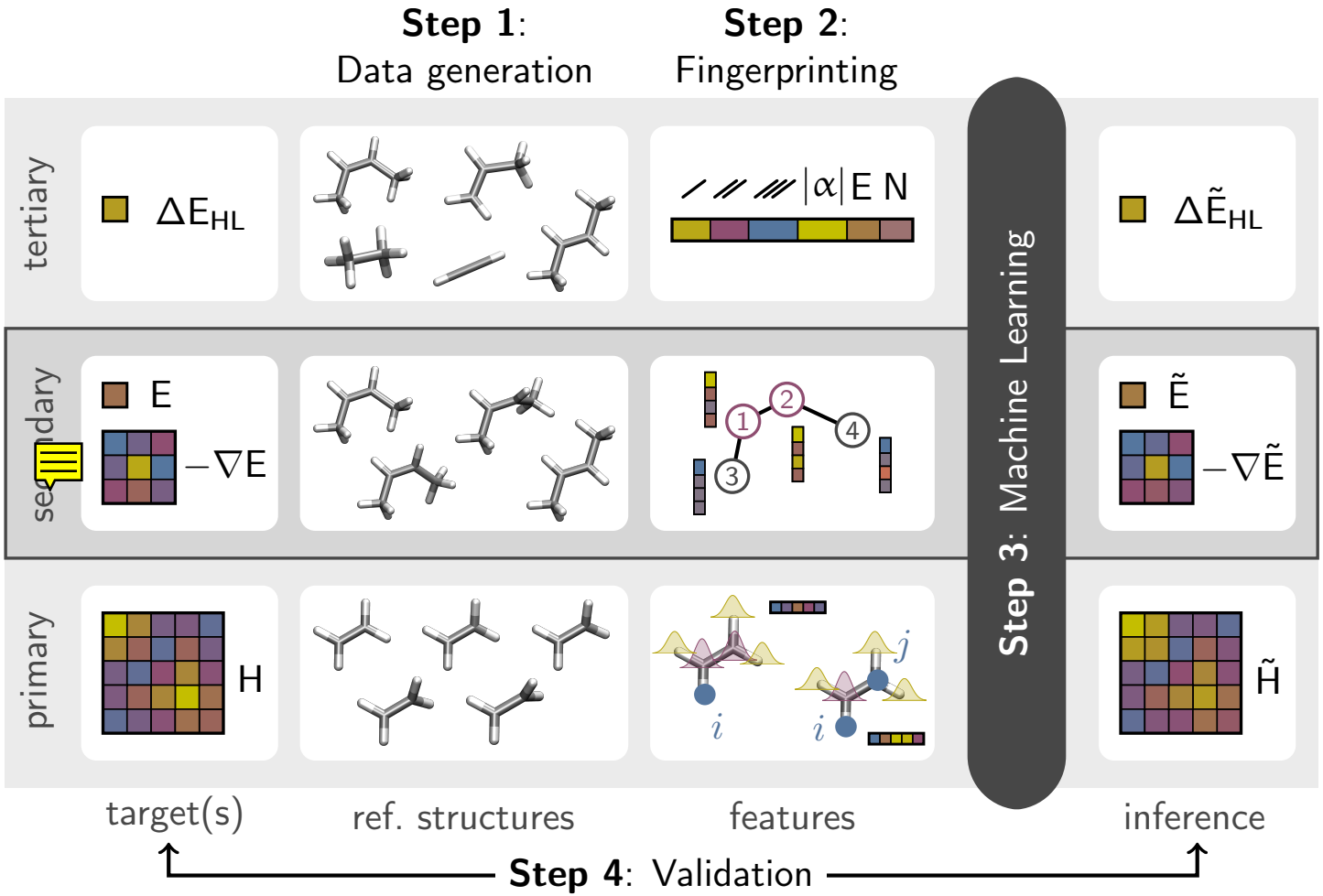
1) Molecular Property Prediction

Typical Targets/Properties



A: Molecular Property Prediction

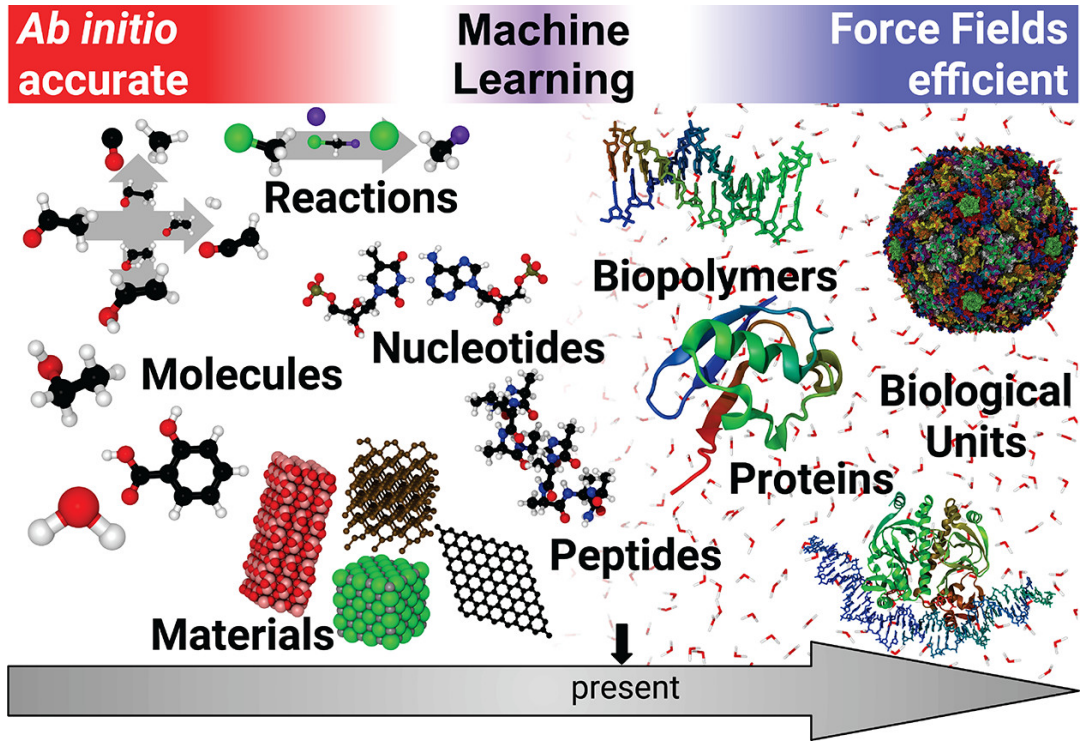
ML Targets



1) Molecular Property Prediction

Machine Learning Force Fields – Motivation

Accurate *ab initio* methods are computationally demanding and can only be used to study small systems in gas phase or regular periodic materials. Larger molecules in solution, such as proteins, are typically modeled by force fields, empirical functions that trade accuracy for computational efficiency. Machine learning methods are closing this gap and allow to study increasingly large chemical systems at *ab initio* accuracy with force field efficiency.



O. T. Unke *et al.*, 10.1021/acs.chemrev.0c01111

1) Molecular Property Prediction

Local Machine Learning Force Fields

Core Idea

- Total energy expressed as sum of atomic contributions:

$$E = \sum_i E_i(\mathbf{G}_i)$$

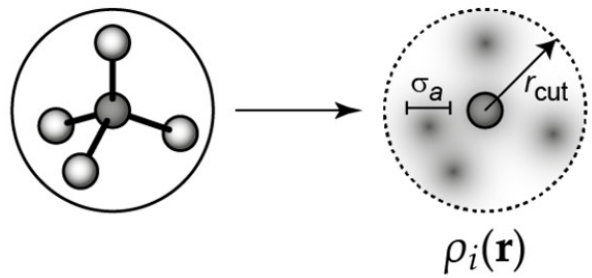
- Each atomic environment is encoded by rotationally/perm.-invariant descriptors.
- Forces obtained by differentiating local energy contributions.

Advantages

- Linear scaling with system size.
- Highly transferable and size-extensive.
- Efficient for molecular dynamics of large systems.
- Strong performance with equivariant neural networks.

Limitations

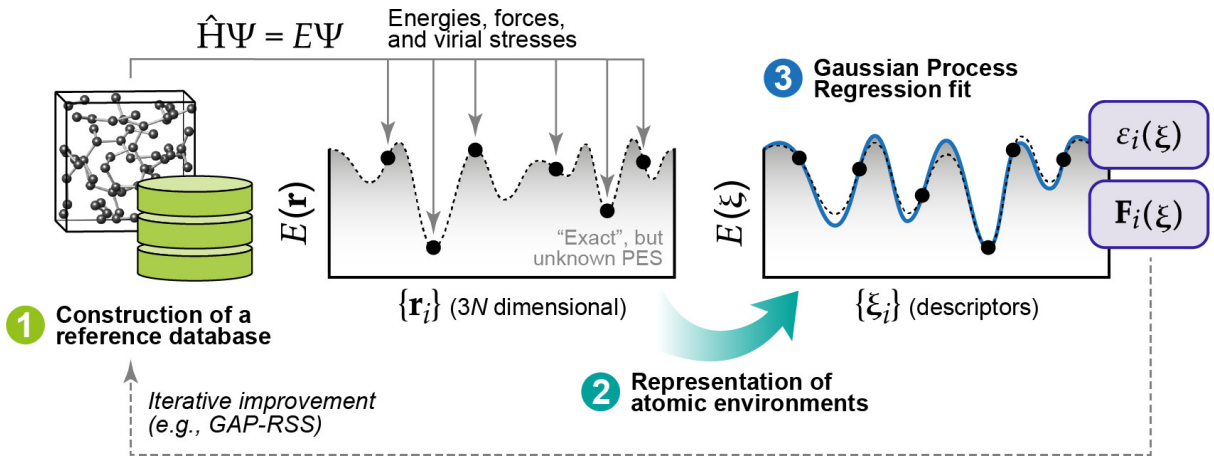
- Implicit or explicit cutoffs truncate long-range physics.
- Message passing only propagates information locally.
- Corrections often needed for electrostatics/dispersion.



V. Deringer *et al.*,
10.1021/acs.chemrev.1c00022

1) Molecular Property Prediction

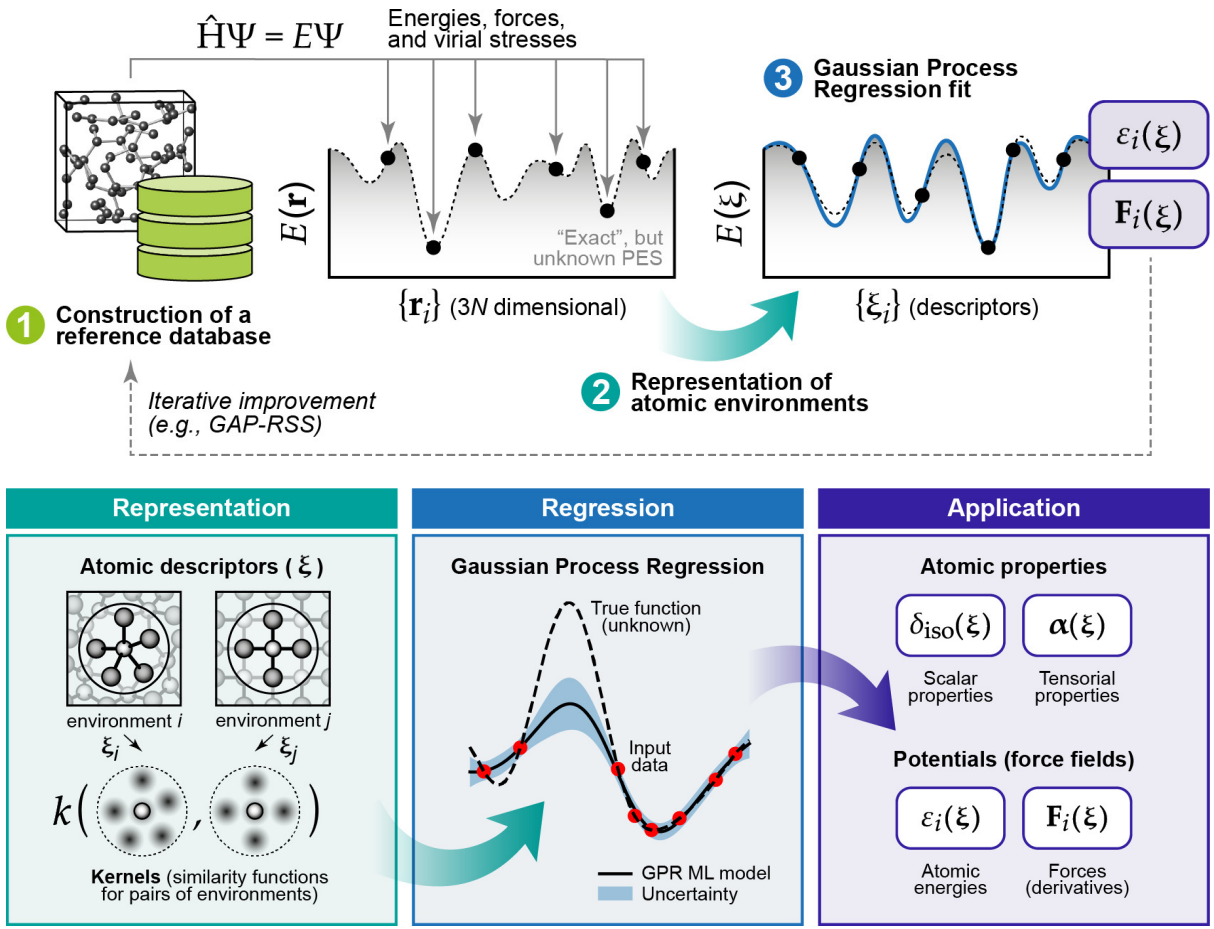
Local Machine Learning Force Fields



V. Deringer *et al.*, 10.1021/acs.chemrev.1c00022

1) Molecular Property Prediction

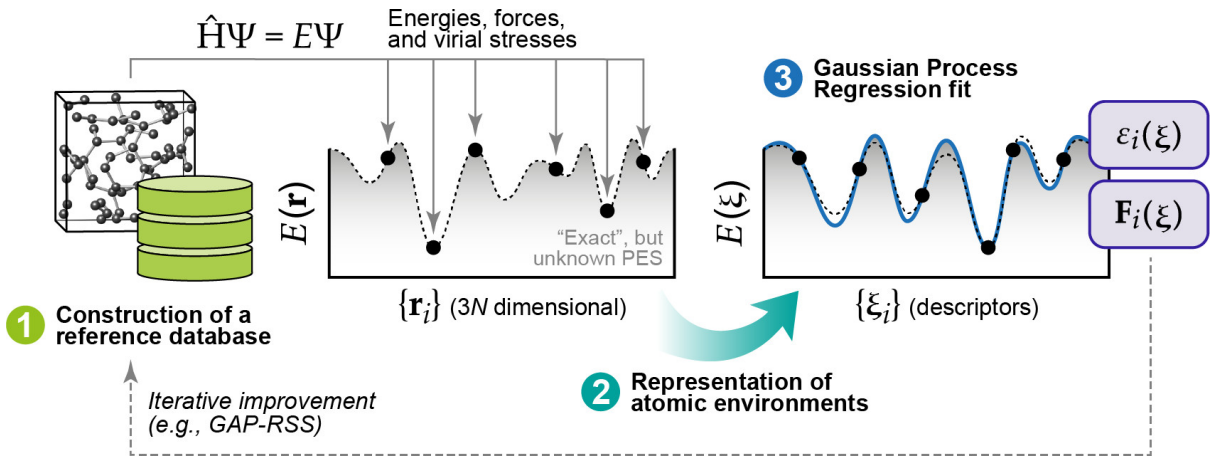
Local Machine Learning Force Fields

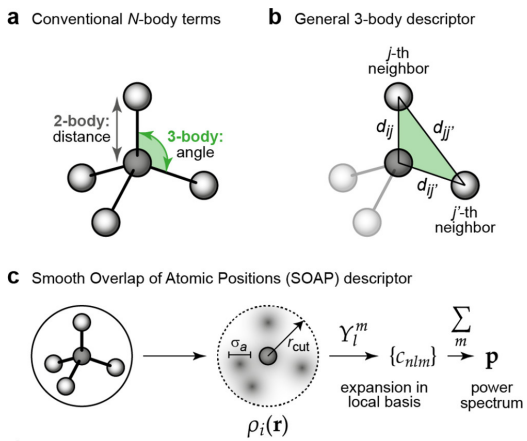


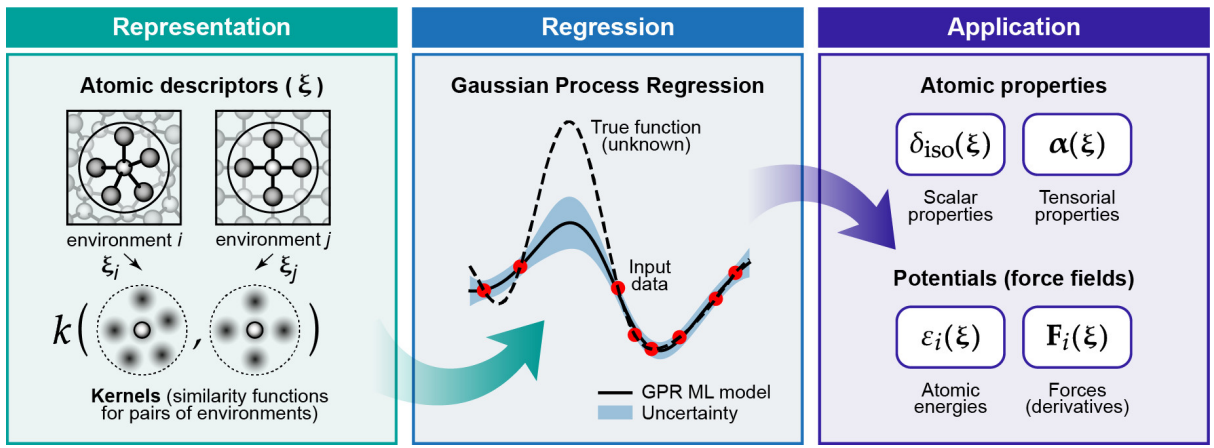
V. Deringer *et al.*, 10.1021/acs.chemrev.1c00022

1) Molecular Property Prediction

Local Machine Learning Force Fields







V. Deringer *et al.*, 10.1021/acs.chemrev.1c00022

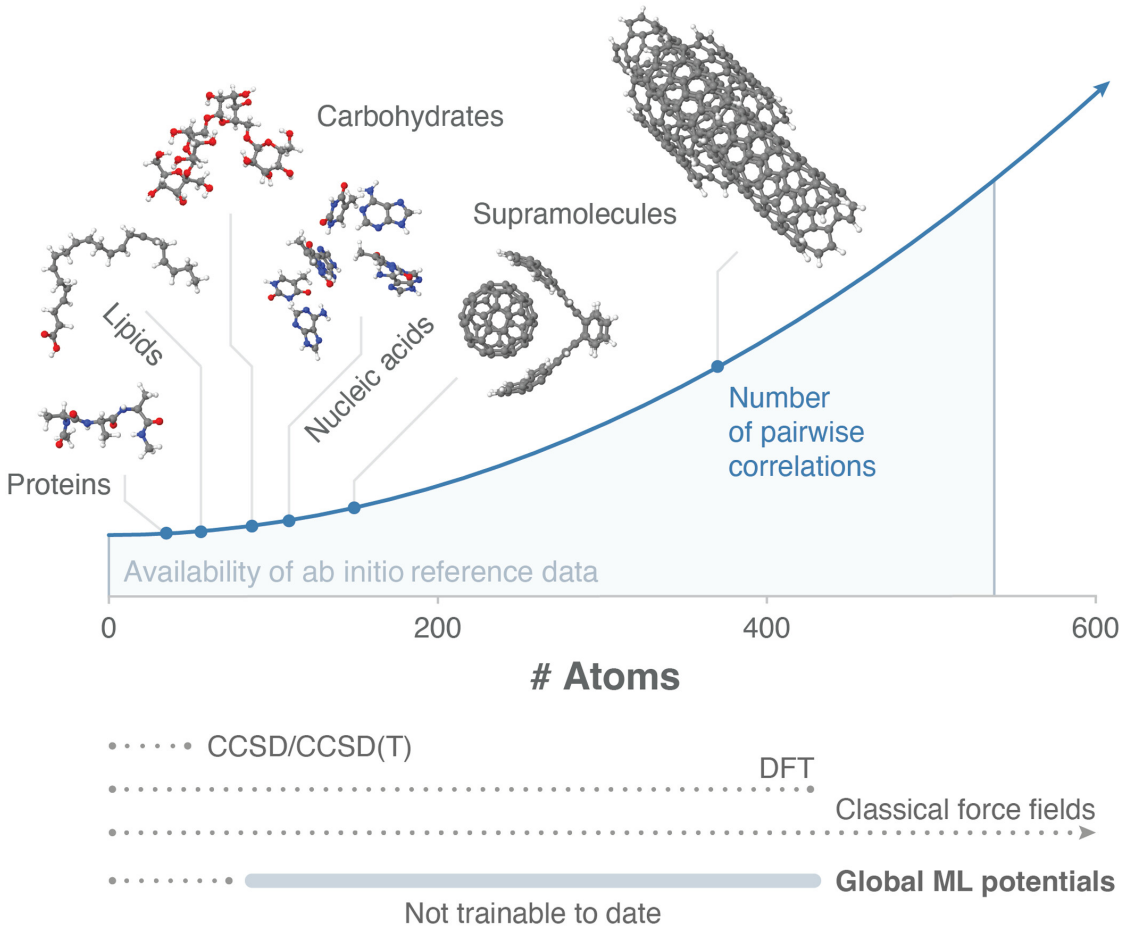
1) Molecular Property Prediction

Global Machine Learning Force Fields

Global MLFF Models

- Capture *all* interaction scales without locality assumptions.
- Avoid information loss caused by truncated long-range interactions.
- Conceptually closest to fully *ab initio* force predictions.

Challenge: quadratic number of atom-atom interactions imposing severe computational limits (computational cost scales approximately as $\mathcal{O}(N^2)$); Current global MLFFs limited to only a few dozen atoms, despite availability of larger *ab initio* datasets.



S. Chmiela *et al.*, 10.1126/sciadv.adf0873

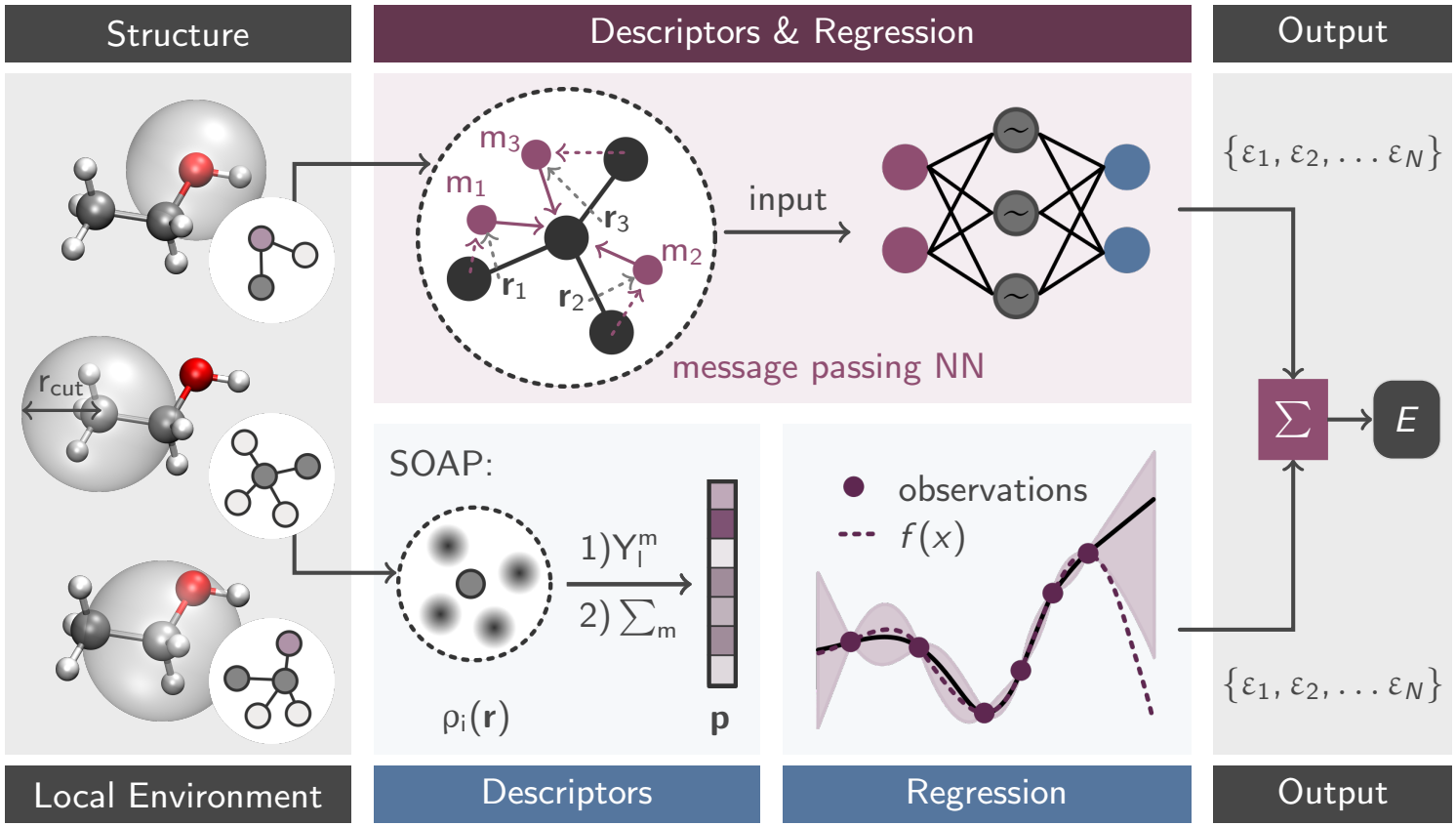
1) Molecular Property Prediction

Local vs. Global Descriptors

Aspect	**Global Descriptors (Kernels)**	**Atomwise Descriptors (NNs)**
Representation	Whole-structure descriptor	Local atomic environments
Scalability	$\mathcal{O}(N^2)$	$\mathcal{O}(N)$ (with atoms N)
Transferability	Limited	High (size-extensive)
Energy Model	$E = \sum_i \alpha_i k(\mathbf{R}, \mathbf{R}_i)$	$E = \sum_i E_i(\mathbf{G}_i)$
Force Computation	Analytic kernel derivatives	Backprop through NN
Data Efficiency	Very high	Moderate to high

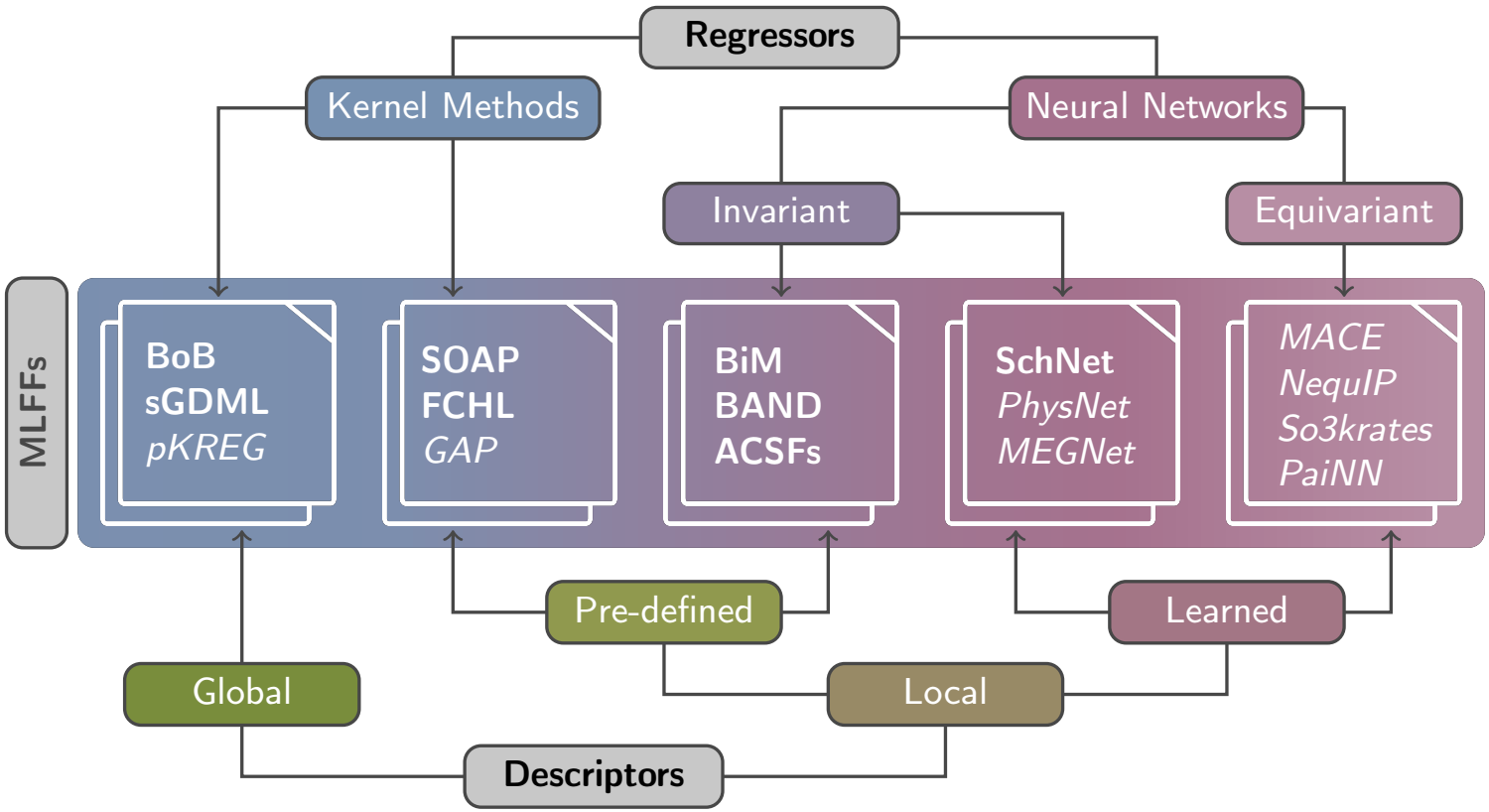
1) Molecular Property Prediction

Local vs. Global Descriptors



1) Molecular Property Prediction

Classification of ML Force Fields



1. Molecular property prediction
2. **Molecule generation and optimization**
3. Reaction prediction and synthesis planning
4. Datasets for Molecule Discovery

2) Molecular Structure Generation

Overview

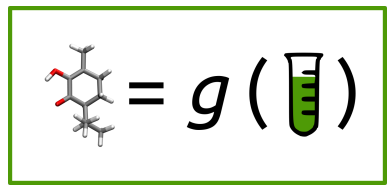
Core Idea: Design molecules with target properties (note: properties are determined by the 3D arrangement of atoms)

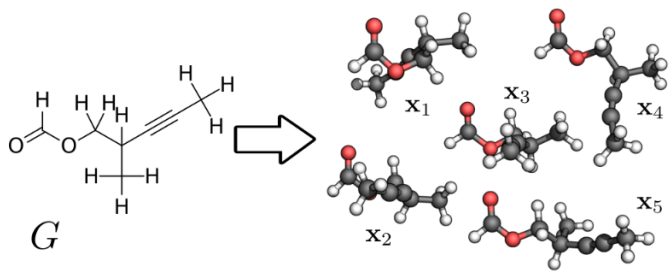
Goals:

- provide plausible conformations for simulations, screening, and ML models
- generate new molecules or conformers/isomers of a structure that satisfy desired properties or functions
- iteratively refine molecules to enhance target properties while maintaining chemical validity

Key Challenges:

- Ensure chemically valid and stereochemically correct molecules
- Explore vast chemical space efficiently
- Produce structures suitable for downstream evaluation (simulations or ML property predictions)





G. Simm & J. M. Hernández-Lobato, 10.5555/3524938.3525768

2) Molecular Structure Generation

Methods

Domain-Knowledge-Based

- Distance Geometry (DG), stochastic embedding
- ETKDG: experimental torsion rules + crystallographic priors

+ robust, fast, chemically informed

− struggles with complex/polycyclic systems; limited generality

Machine Learning-Based

- Predict 3D coordinates or distance matrices
- techniques: variational autoencoders, reinforcement learning, genetic algorithms

+ captures complex geometry; scalable with data

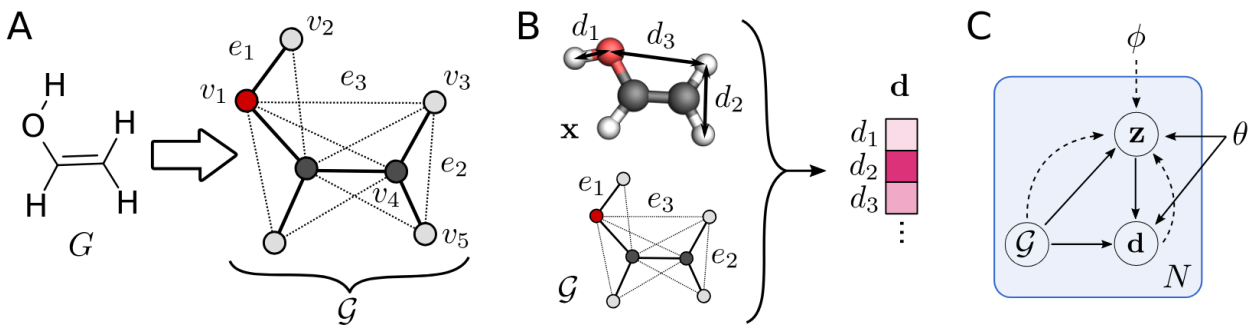
− element range, stereochemistry, generalizability

Grammar-Based

- DeepSMILES, SELFIES, Group SELFIES etc.
- Generative models output sequences decoded to molecular graphs (guarantees syntactic validity)

+ simple, compatible with NLP models

− 3D geometry must still be embedded afterwards



G. Simm & J. M. Hernández-Lobato, 10.5555/3524938.3525768

1. Molecular property prediction
2. Molecule generation and optimization
3. **Reaction prediction and synthesis planning**
4. Datasets for Molecule Discovery

3) Reaction & Synthesis Planning

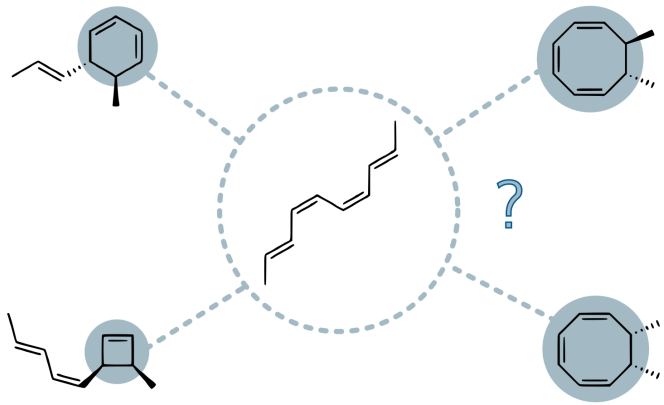
Overview

Goal: Predict Reactions and Plan Synthetic Routes

- Predict feasible chemical reactions given reactants and conditions.
- Design multi-step synthetic pathways to obtain target molecules.
- Optimize reaction conditions and route efficiency (yield, cost, safety).
- Aid chemists in discovering novel synthetic strategies and exploring chemical space systematically.

Why is this needed?

- astronomically large space of possible reactions and synthetic routes
- manual planning is time-consuming and error-prone
- accelerate synthesis planning and support decision-making
- enable (semi-)automated lab workflows for molecule discovery



C. Coley *et al.*, 10.1021/acs.jctc.5c01161

1. Molecular property prediction
2. Molecule generation and optimization
3. Reaction prediction and synthesis planning
4. **Datasets for Molecule Discovery**

Dataset 1

QM9

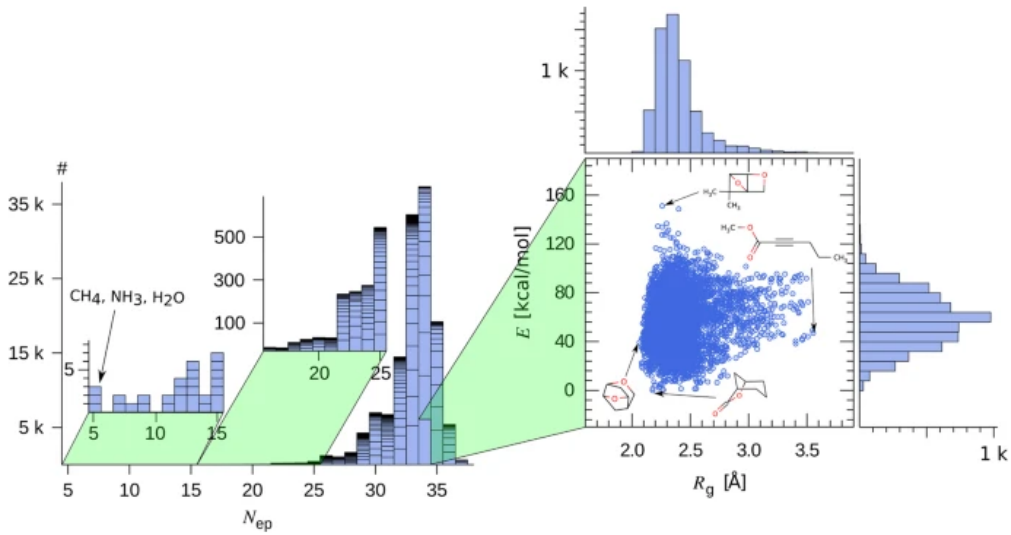
Data Descriptor | Open access | Published: 05 August 2014

Quantum chemistry structures and properties of 134 kilo molecules

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp & O. Anatole von Lilienfeld

Scientific Data **1**, Article number: 140022 (2014) | Cite this article

90k Accesses | **1500** Citations | **47** Altmetric | Metrics



R. Ramakrishnan *et al.*, 10.1038/sdata.2014.22

Dataset 2

QM7-x

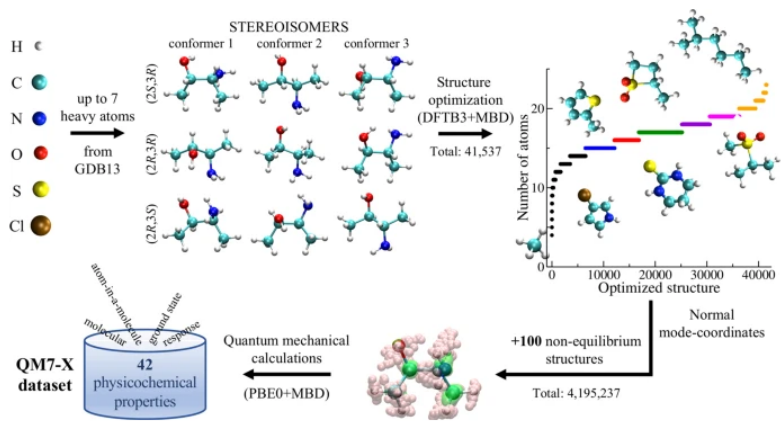
Data Descriptor | Open access | Published: 02 February 2021

QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules

Johannes Hoja, Leonardo Medrano Sandonas, Brian G. Ernst, Alvaro Vazquez-Mayagoitia, Robert A. DiStasio Jr. & Alexandre Tkatchenko

Scientific Data **8**, Article number: 43 (2021) | Cite this article

17k Accesses | **130** Citations | **23** Altmetric | Metrics



J. Hoja *et al.*, 10.1038/s41597-021-00812-2

Dataset 3

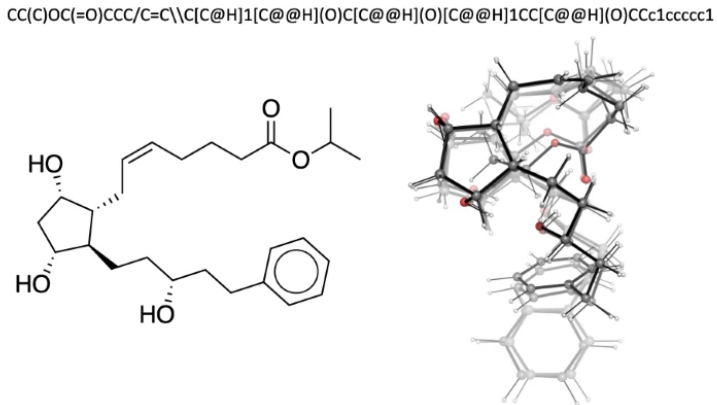
GEOM

Data Descriptor | Open access | Published: 21 April 2022

GEOM, energy-annotated molecular conformations for property prediction and molecular generation

Simon Axelrod & Rafael Gómez-Bombarelli

Scientific Data **9**, Article number: 185 (2022) | Cite this article



Molecular representations of the latanoprost molecule. *top* SMILES string. *left* Stereochemical formula with edge features, including wedges for in- and out-of-plane bonds, and a double line for *cis* isomerism. *right* Overlay of conformers. Higher transparency corresponds to lower statistical weight.

S. Axelrod & R. Gómez-Bombarelli, 10.1038/s41597-022-01288-4

Dataset 4

WS22

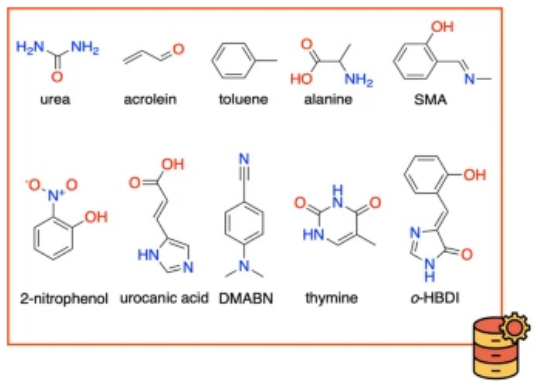
Data Descriptor | Open access | Published: 15 February 2023

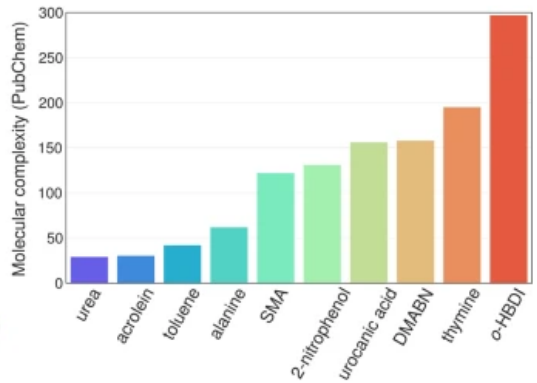
WS22 database, Wigner Sampling and geometry interpolation for configurationally diverse molecular datasets

Max Pinheiro Jr, Shuang Zhang, Pavlo O. Dral & Mario Barbatti

Scientific Data **10**, Article number: 95 (2023) | Cite this article

6076 Accesses | **21** Citations | **19** Altmetric | Metrics





M. Pinheiro Jr *et al.*, 10.1038/s41597-023-01998-3

Dataset 5

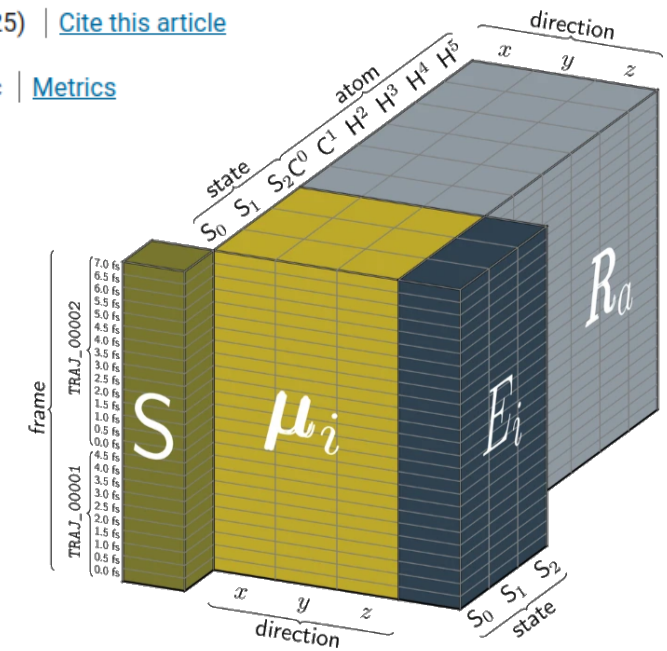
shnitsel-data

Data Descriptor | Open access | Published: 26 July 2025

Surface Hopping Nested Instances Training Set for Excited-state Learning

Robin Curth, Theodor E. Röhrkasten, Carolin Müller & Julia Westermayr

Scientific Data **12**, Article number: 1300 (2025) | Cite this article



R. Curth *et al.*, 10.1038/s41597-025-05443-5

Dataset 6

Deep4Chem

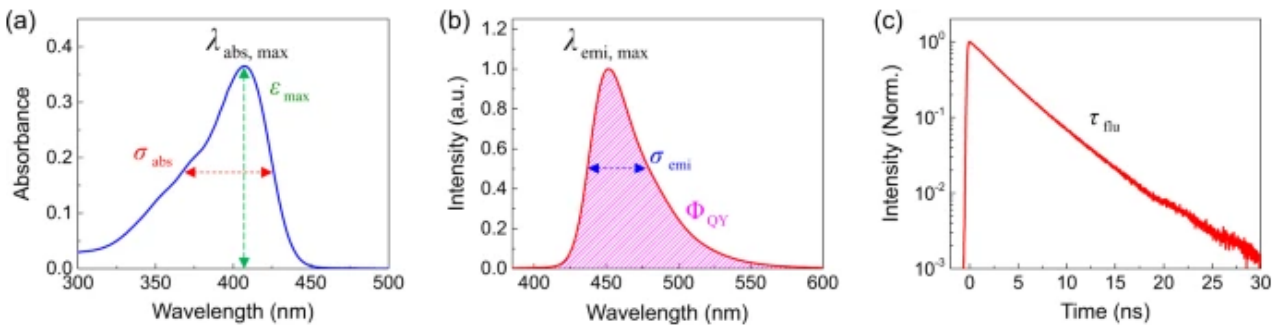
Data Descriptor | Open access | Published: 08 September 2020

Experimental database of optical properties of organic compounds

Joonyoung F. Joung, Minhi Han, Minseok Jeong & Sungnam Park

Scientific Data **7**, Article number: 295 (2020) | Cite this article

27k Accesses | **84** Citations | **10** Altmetric | Metrics



J. F. Joung *et al.*, 10.1038/s41597-020-00634-8

Dataset 7

xxMD

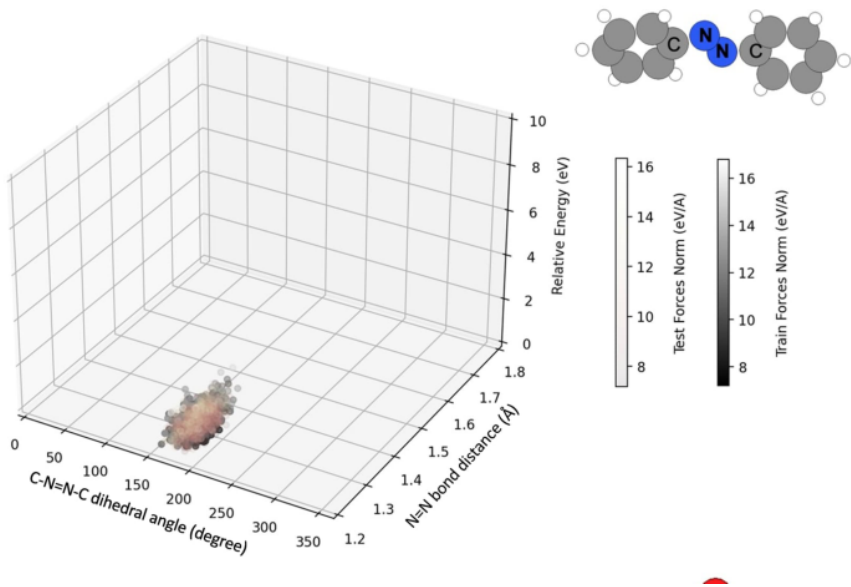
Data Descriptor | Open access | Published: 20 February 2024

Beyond MD17: the reactive xxMD dataset

Zihan Pengmei, Junyu Liu & Yinan Shu

Scientific Data **11**, Article number: 222 (2024) | Cite this article

4563 Accesses | **8** Citations | **1** Altmetric | Metrics



Z. Pengmei *et al.*, 10.1038/s41597-024-03019-3

Dataset 8

COMPAS-2

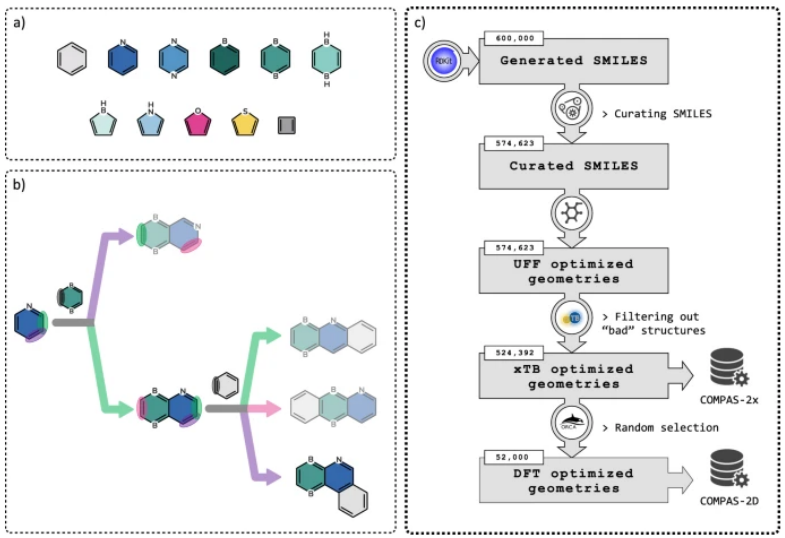
Data Descriptor | Open access | Published: 19 January 2024

COMPAS-2: a dataset of *cata*-condensed hetero-polycyclic aromatic systems

Eduardo Mayo Yanes, Sabyasachi Chakraborty & Renana Gershoni-Poranne

Scientific Data **11**, Article number: 97 (2024) | Cite this article

4299 Accesses | **16** Citations | **9** Altmetric | Metrics



E. M. Yanes *et al.*, 10.1038/s41597-024-02927-8

Dataset 9

Open Reaction Database (ORD)

Journal of the American Chemical Society > Vol 143/Issue 45 > Article

Open Access

Cite Share Jump to Expand

PERSPECTIVE | November 2, 2021

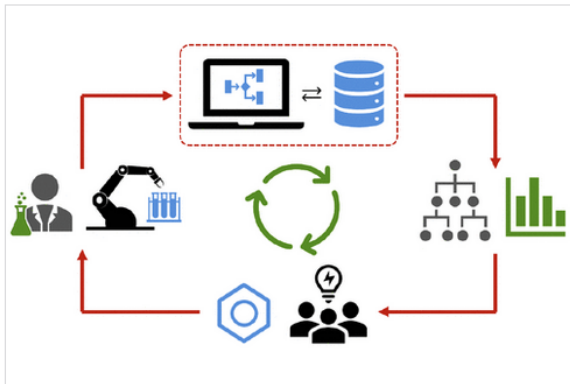
The Open Reaction Database

Steven M. Kearnes*, Michael R. Maser, Michael Wleklinski, Anton Kast, Abigail G. Doyle, Spencer D. Dreher, Joel M. Hawkins, Klavs F. Jensen, and Connor W. Coley*

Open PDF S·F·X

Abstract

Chemical reaction data in journal articles, patents, and even electronic laboratory notebooks are currently stored in various formats, often unstructured, which presents a significant barrier to downstream applications, including the training of machine-learning models. We present the Open Reaction Database (ORD), an open-access schema and infrastructure for structuring and sharing organic reaction data, including a centralized data repository. The ORD schema supports conventional and emerging technologies, from benchtop reactions to automated high-throughput experiments and flow chemistry. The data, schema, supporting code, and web-based user interfaces are all publicly available on GitHub. Our vision is that a consistent data representation and infrastructure to support data sharing will enable downstream applications that will greatly improve the state of the art with respect to computer-aided synthesis planning, reaction prediction, and other predictive chemistry tasks.



S. M. Kearnes *et al.*, 10.1021/jacs.1c09820

Dataset 10

HLM dataset (ADMET)

Chemical Research in Toxicology > Vol 35/Issue 9 > Article

Subscribed

Cite Share Jump to Expand

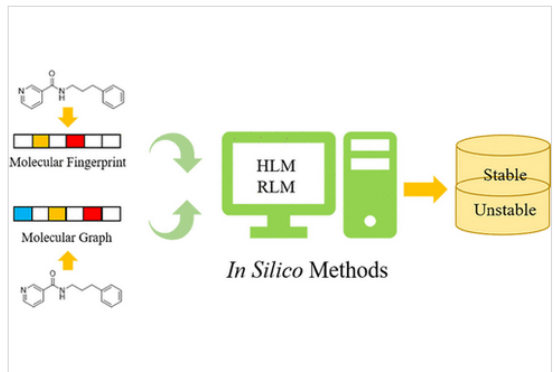
ARTICLE | September 2, 2022

In Silico Prediction of Human and Rat Liver Microsomal Stability via Machine Learning Methods

Longqiang Li, Zhou Lu, Guixia Liu, Yun Tang, and Weihua Li*

Open PDF | Supporting Information (2) | S·F·X

Abstract



Liver microsomal stability is an important property considered for the screening of drug candidates in the early stage of drug development. Determination of hepatic metabolic stability can be performed by an in vitro assay, but it requires quite a few resources and time. In recent years, machine learning methods have made much progress. Therefore, development of computational models to predict liver microsomal stability is highly desirable in the drug discovery process. In this study, the in silico classification models for the prediction of the metabolic stability of compounds in rat and human liver microsomes were constructed by the conventional machine learning and deep learning methods. The performance of the models was evaluated using the test and external sets. For the rat liver microsomes (RLM) stability, the best model yielded the AUC values of 0.84 and 0.71 on the test and external validation sets, respectively. For the human liver microsome (HLM) stability, the best model exhibited the AUC values of 0.86 and 0.77 on the test and external validation sets, respectively. In addition, several important substructure fragments were detected using information gain and frequency substructure analysis methods. The applicability domain of the models was defined using the Euclidean distance-based method. We anticipate that our results would be helpful for the prediction of liver microsomal stability of compounds in the early stage of drug discovery.

L. Li *et al.*, 10.1021/acs.chemrestox.2c00207

Dataset 11

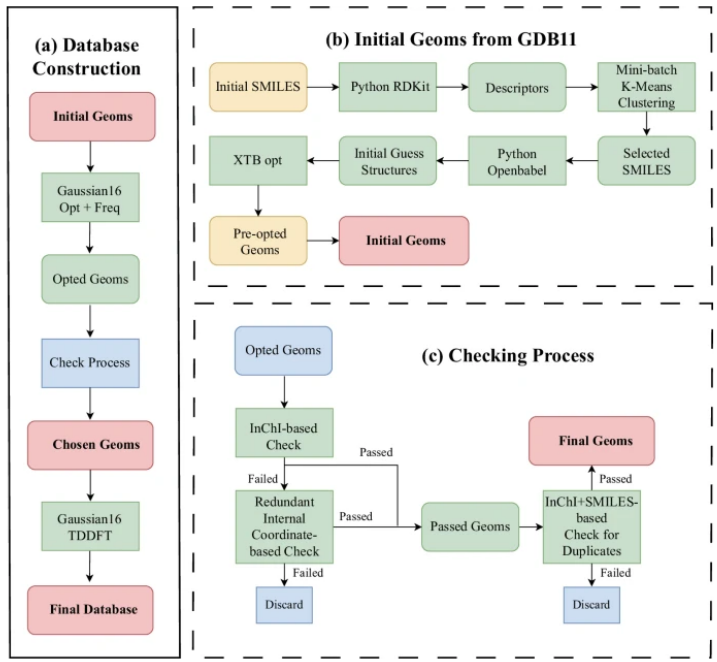
QCDGE

Data Descriptor | Open access | Published: 29 August 2024

Quantum Chemistry Dataset with Ground- and Excited-state Properties of 450 Kilo Molecules

Yifei Zhu, Mengge Li, Chao Xu & Zhenggang Lan

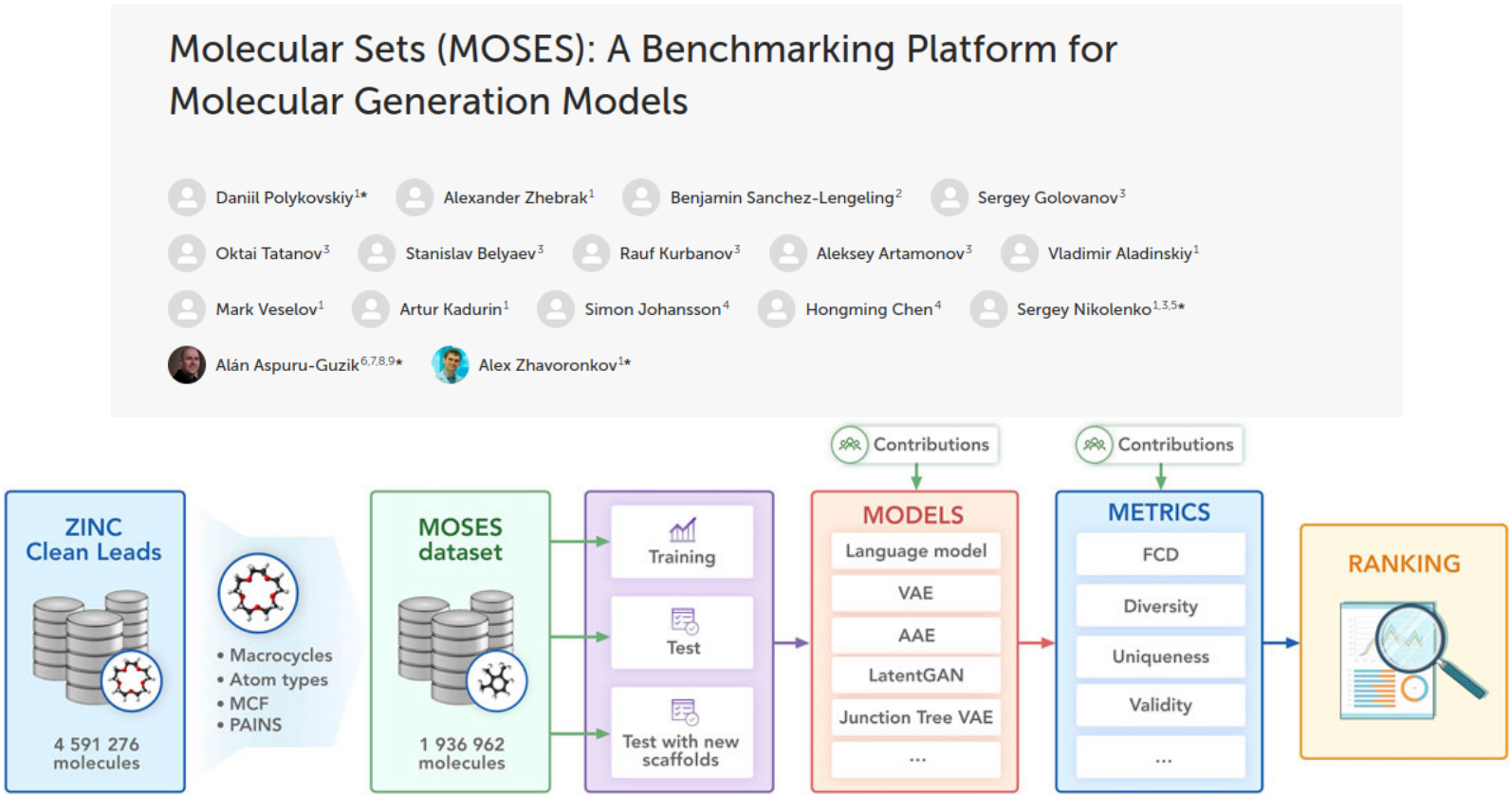
Scientific Data **11**, Article number: 948 (2024) | Cite this article



Y. Zhu *et al.*, 10.1038/s41597-024-03788-x

Dataset 12

MOSES



D. Polykovskiy *et al.*, 10.3389/fphar.2020.565644

Dataset

Assignments

QM9

WS22

xxMD

HLM

QM7-x

shnitsel-data

COMPAS-2

QCDGE

GEOM

Deep4Chem

ORD

MOSES

Student Task: Dataset Presentation

Objective: Prepare a **7-minute talk** on your chosen dataset (until December 8 or 15).

Your presentation should cover:

1. **Dataset Overview**
   - What information is stored (molecules, reactions, properties, etc.)
   - Level of data: experimental or computational
   - Key metadata (units, conditions, identifiers, etc.)
   - visualize key aspects of the dataset (e.g. using histograms, PCA plots, etc.)
2. **Working with the Dataset**
   - Example code snippets or live demonstration e.g. using a Jupyter notebook (loading, preprocessing, visualization)
   - Any special tools or libraries required
3. **Applications**
   - assign the dataset to one of the main application blocks: 1) Molecular property prediction, 2) Molecular structure generation, 3) Synthesis planning
   - show examples from the literature how they employed the dataset and what was achieved
   - Propose **one original idea** for a project:
     - Define a key goal
     - Outline a tentative plan to achieve it

Hands-On

Objective: Supervised classification of dyes into **cyanine** and **acridine** classes.

Dataset:

- 40 dye molecules with known class labels
- One molecule has a missing class label (to be predicted)
- Features: structural and/or electronic descriptors

Task Steps:

1. Create two scalar (single-value) invariant features from the dataset
2. Apply classification methods of your choice:
   - Examples: k-Nearest Neighbors, Logistic Regression, Random Forest
3. Train models to classify the dyes and predict the missing label
4. Evaluate performance:
   - Maximum accuracy achievable
   - Identify which features and method yield best performance
5. Send your final results to Prof. Müller *via* email (until 02nd of December)