Pacific Northwest
NATIONAL LABORATORY

Data-driven Scientific Discovery

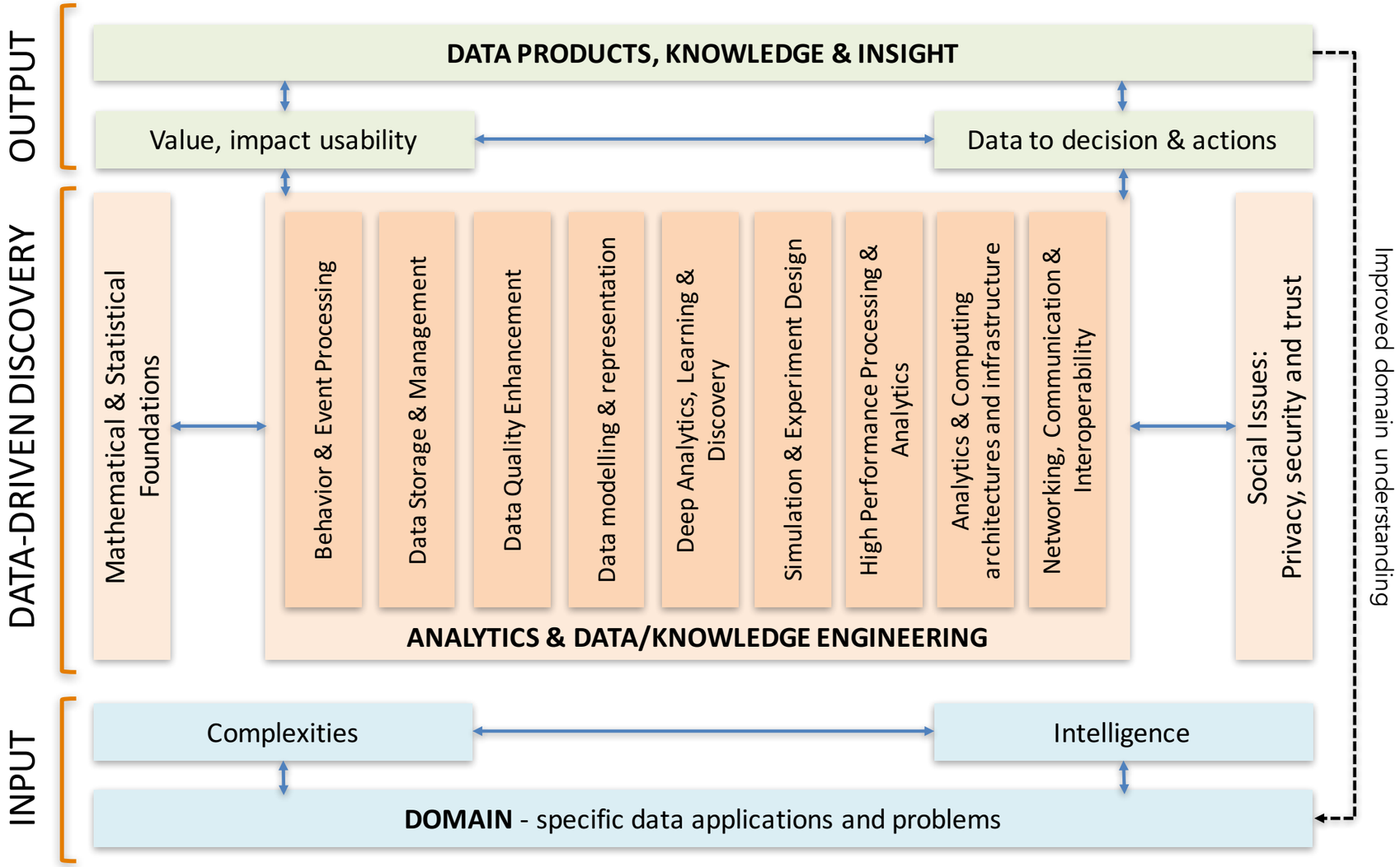
July 12, 2021

Robert Rallo
Director
Advanced Computing, Mathematics, and Data Division

U.S. DEPARTMENT OF ENERGY BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

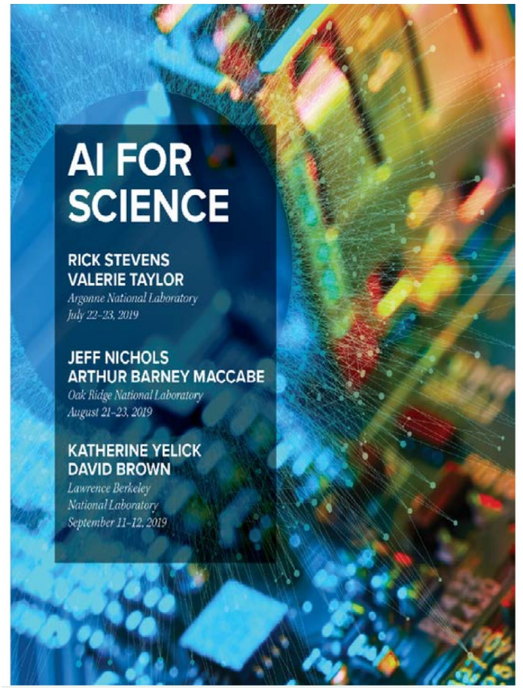
Data Science



Adapted from Cao, L. (2017). "Data Science: A Comprehensive Overview." ACM Comput. Surv. **50**(3): Article 43.

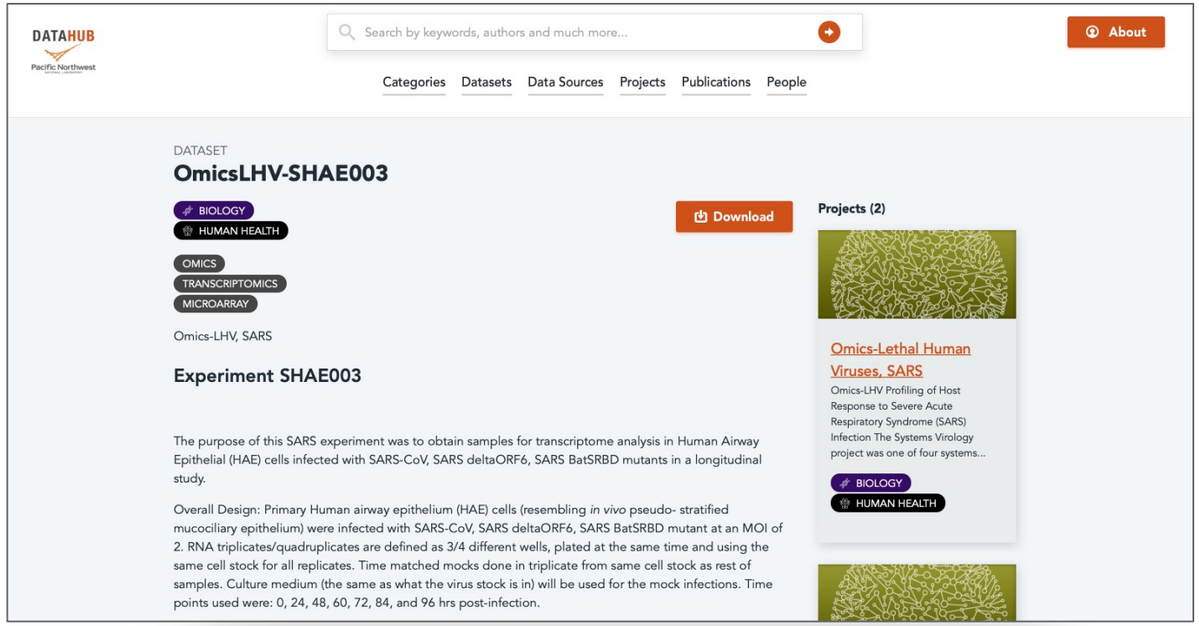
Scientific Machine Learning

- Machine Learning algorithms should be:
  - **Domain-aware**
  - Interpretable
  - Robust
  - Data-intensive
  - Scalable, deployable
- Machine Learning must contribute to:
  - Enhance current modeling and simulation
  - Automation and decision support



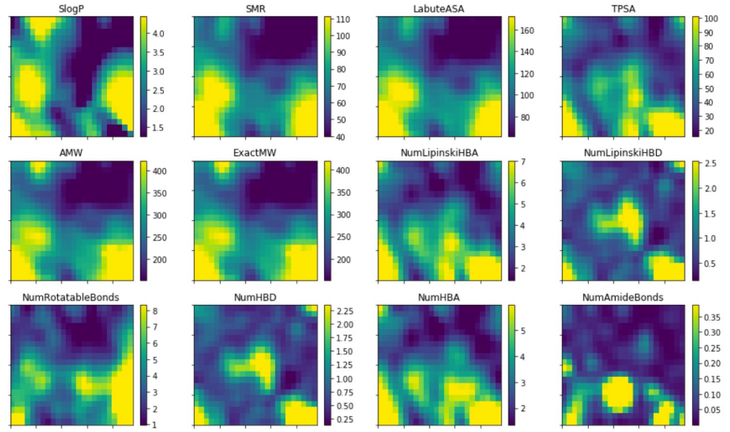
Data

- Data products
  - Beyond raw data
  - Models, workflows
- Data governance and data stewardship
  - Define process and procedures
  - Ensure compliance and security
  - Manages data, tools and processes
  - Comprehensive policies
- Data quality
  - Metrics
- Privacy preservation
  - Privacy-preserving analytics
  - Privacy-preserving AI/ML

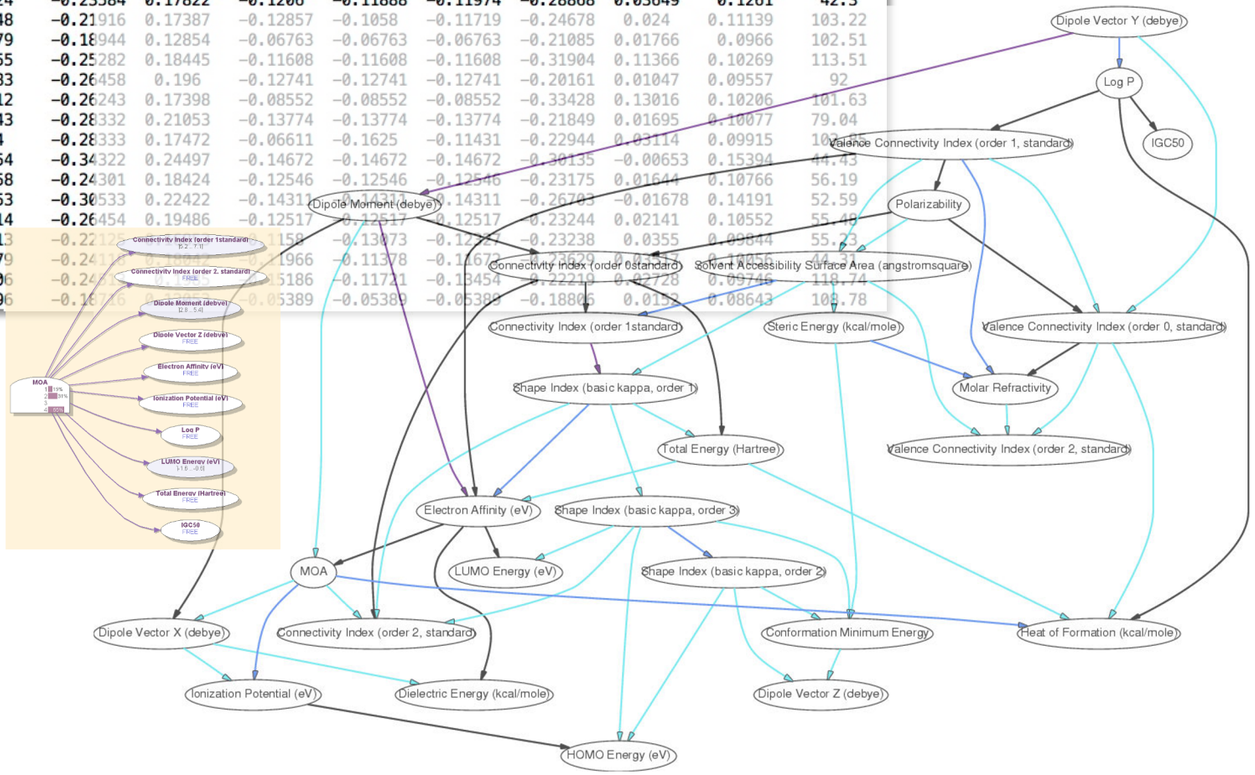


Data Representation

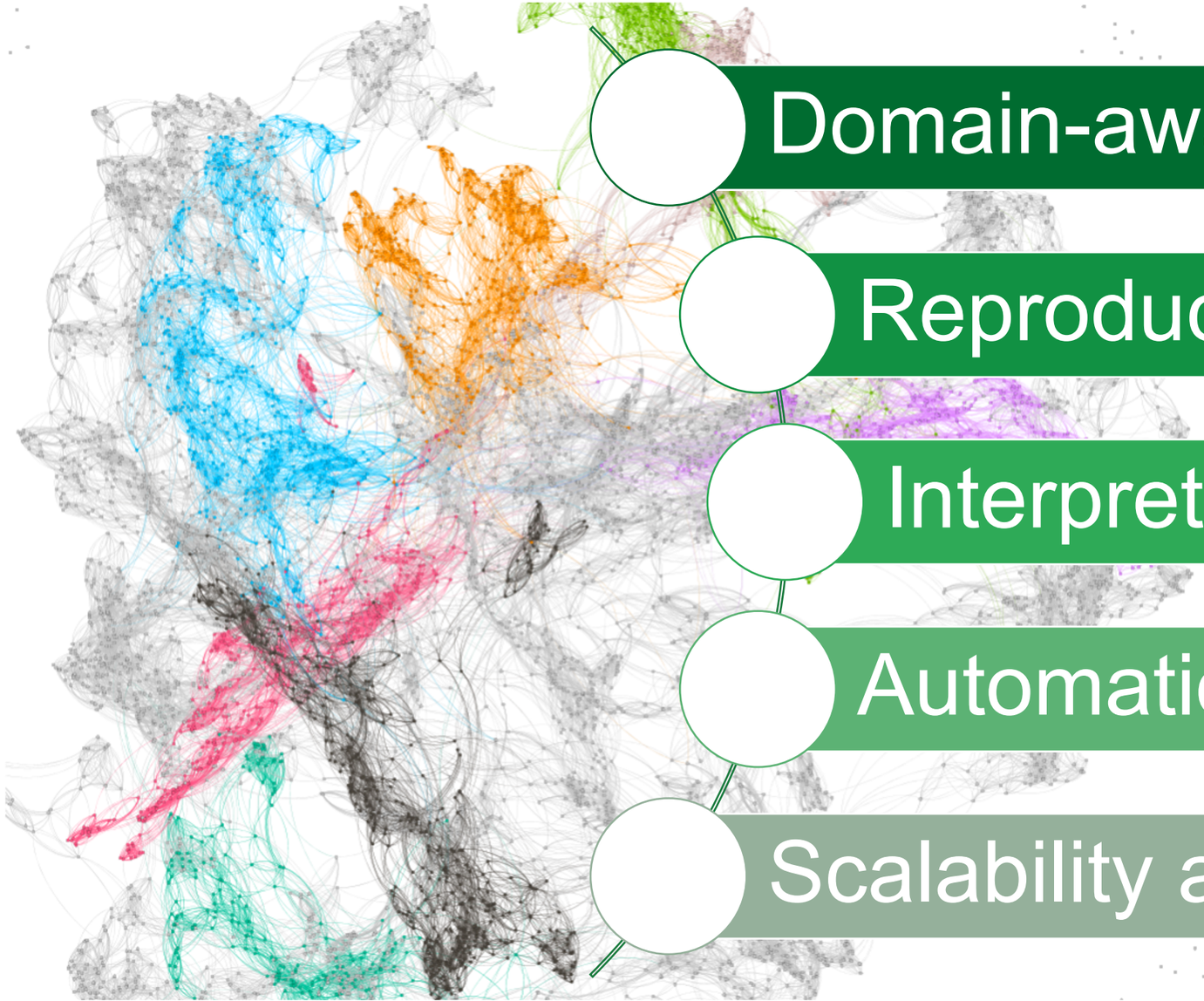
- Structured vs unstructured data
  - e.g., knowledge extraction from literature sources
- Efficient data structures
  - graphs
- Feature engineering
  - feature selection
  - feature synthesis
- **Representation learning**
  - interpretability



material	light_tox	dark_tox	HHOMO	LZELEHHO	LUMOA	LUMOB	ALZLUMO	MHOMOA	MLUMOA	QMELECT	Cp
Zn0	6.23	5.8	−0.22819	0.13198	−0.03576	−0.13869	−0.08723	−0.29251	0.04335	0.12458	40.25
Cu0	5.71	4.24	−0.23584	0.17822	−0.1206	−0.11888	−0.11974	−0.28868	0.03649	0.1261	42.3
V203	3.78	3.48	−0.21916	0.17387	−0.12857	−0.1058	−0.11719	−0.24678	0.024	0.11139	103.22
Y203	5.84	5.79	−0.18944	0.12854	−0.06763	−0.06763	−0.06763	−0.21085	0.01766	0.0966	102.51
Bi203	4.02	3.55	−0.25282	0.18445	−0.11608	−0.11608	−0.11608	−0.31904	0.11366	0.10269	113.51
In203	3.48	2.83	−0.26458	0.196	−0.12741	−0.12741	−0.12741	−0.20161	0.01047	0.09557	92
Sb203	3.66	3.12	−0.26243	0.17398	−0.08552	−0.08552	−0.08552	−0.33428	0.13016	0.10206	101.63
Al203	2.75	2.43	−0.28332	0.21053	−0.13774	−0.13774	−0.13774	−0.21849	0.01695	0.10077	79.04
Fe203	2.54	2.4	−0.28333	0.17472	−0.06611	−0.1625	−0.11431	−0.22944	0.03114	0.09915	10
Si02	2.92	2.54	−0.34322	0.24497	−0.14672	−0.14672	−0.14672	−0.30135	−0.00653	0.15394	44.43
Zr02	3.04	2.58	−0.24301	0.18424	−0.12546	−0.12546	−0.12546	−0.23175	0.01644	0.10766	56.19
Sn02	3.24	2.53	−0.30533	0.22422	−0.14311	−0.14311	−0.14311	−0.26703	−0.01678	0.14191	52.59
Ti02	4.68	2.14	−0.26454	0.19486	−0.12517	−0.12517	−0.12517	−0.23244	0.02141	0.10552	55.48
Co0	3.33	3.13	−0.22128		−0.1158	−0.13073	−0.12153	−0.23238	0.0355	0.09844	55.23
Ni0	3.87	3.79	−0.24		−0.11966	−0.11378	−0.11672	−0.25629	0.02425	0.10056	44.23
Cr203	2.06	2.06	−0.24		−0.15186	−0.11721	−0.13454	−0.22219	0.02728	0.09746	118.74
La203	5.56	4.96	−0.18		−0.05389	−0.05389	−0.05389	−0.18806	0.0153	0.08643	109.78



Opportunities for Accelerating Scientific Discovery with AI/ML



- Domain-awareness
- Reproducibility & Robustness
- Interpretability
- Automation
- Scalability and Deployment

Domain-awareness

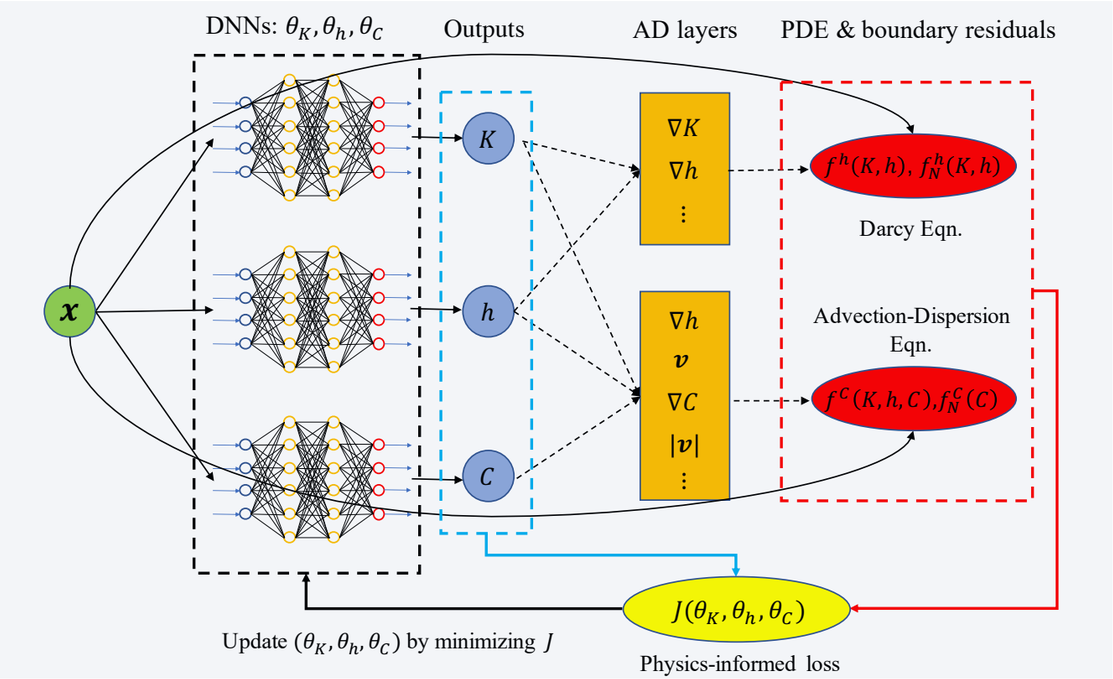
DATA

- Knowledge representations
  - Equations
  - Knowledge graphs
  - Causal models
  - Distributed knowledge models
- Domain-aware features

ALGORITHMS

- Knowledge embedding
  - Constraints
  - Loss function
  - Coupling with simulations
  - Model structure

Coupling limited experimental data with physics models improves accuracy and reduces the risk of model overfitting in subsurface transport modeling.

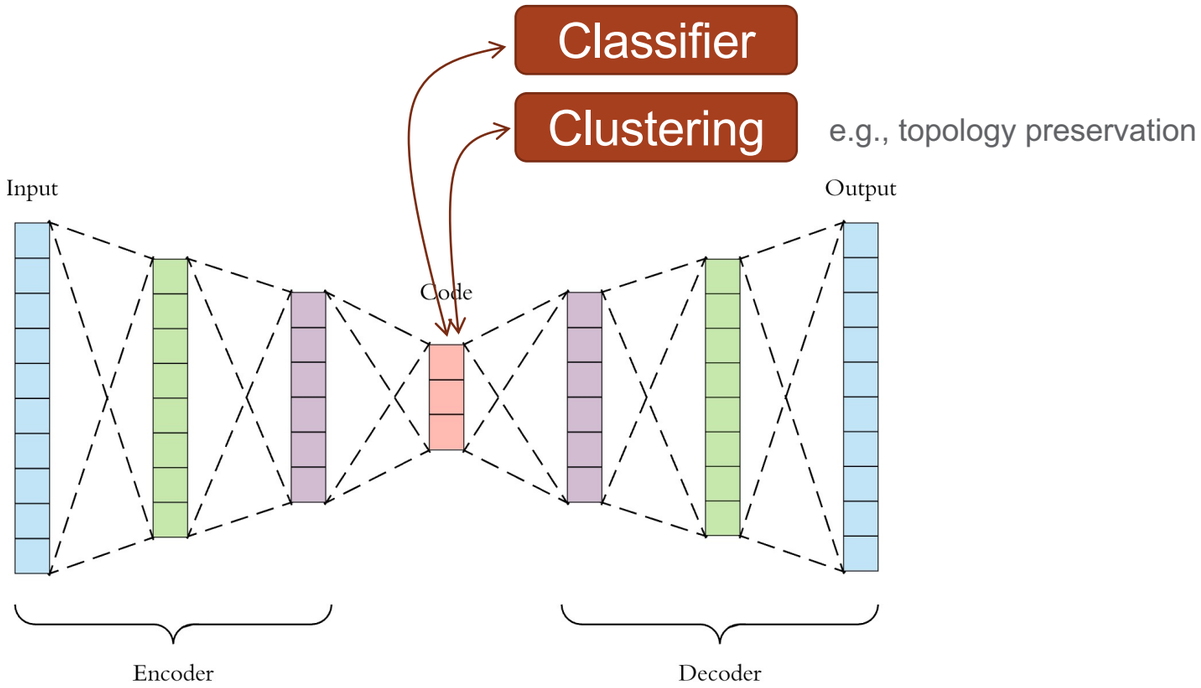


He Q et al.. 2020 Physics-informed neural networks for multi-physics data assimilation with application to subsurface transport. *Advances in Water Resources* (2020) 141.

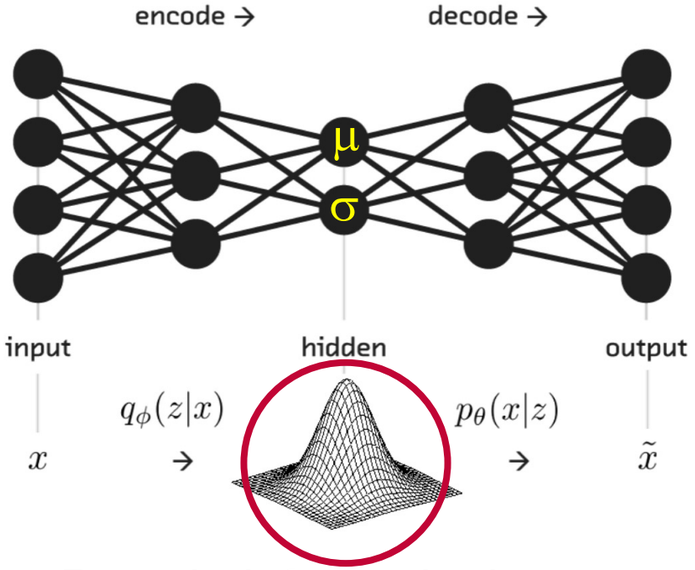
Domain-aware features

- Coupling autoencoders with:
  - Clustering algorithms
  - Classification algorithms

$$\mathcal{L}(x, \hat{x}) + \lambda_1 \sum_i |a_i| + \lambda_2 \mathcal{L}_{\substack{clustering \\ classifier}}$$

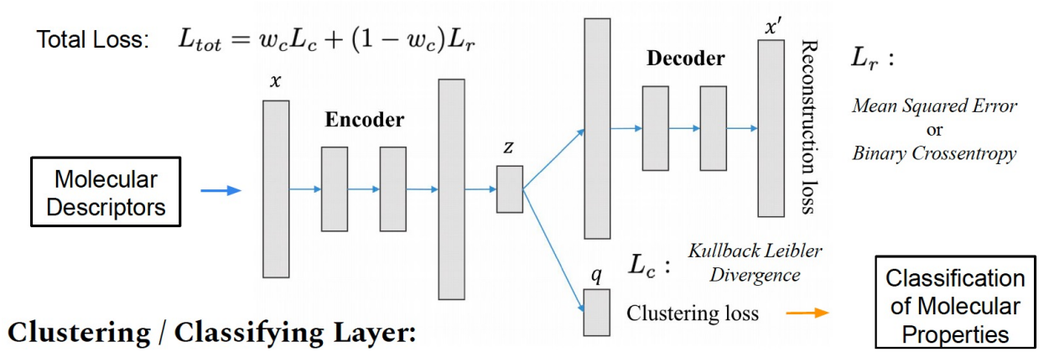


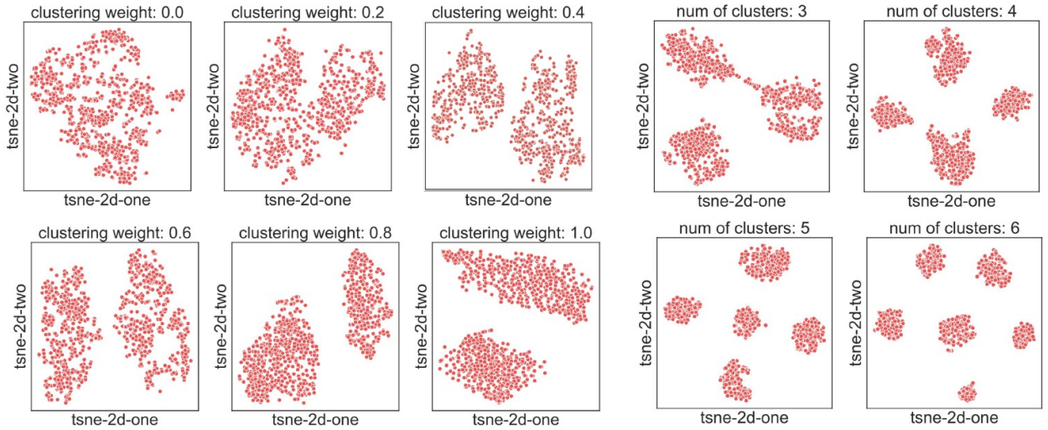
- Variational autoencoders with:
  - Predefined priors



- Domain-informed prior
- KL term in the loss function enforces the desired probability distribution

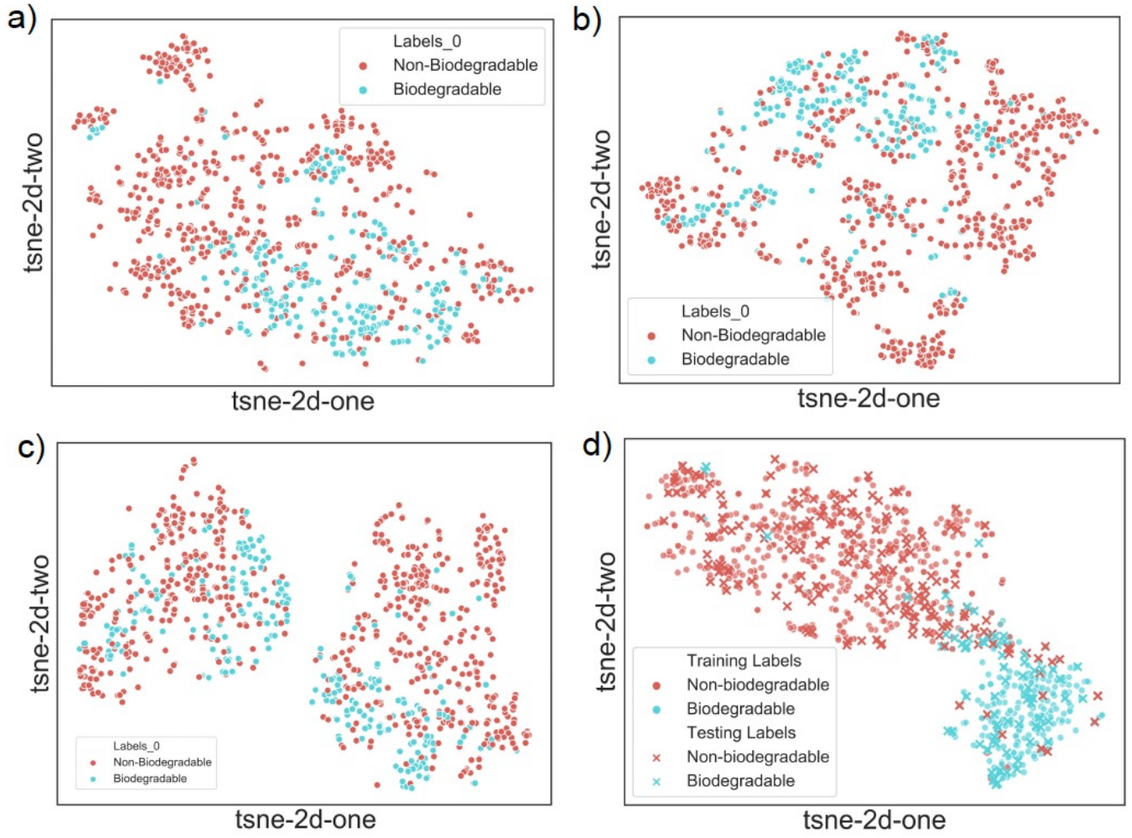
Biodegradability-aware embeddings





The 2-dimensional T-SNE projections of hidden layers, number of clusters = 2

The 2-dimensional T-SNE projections of hidden layers, clustering weights = 0.6

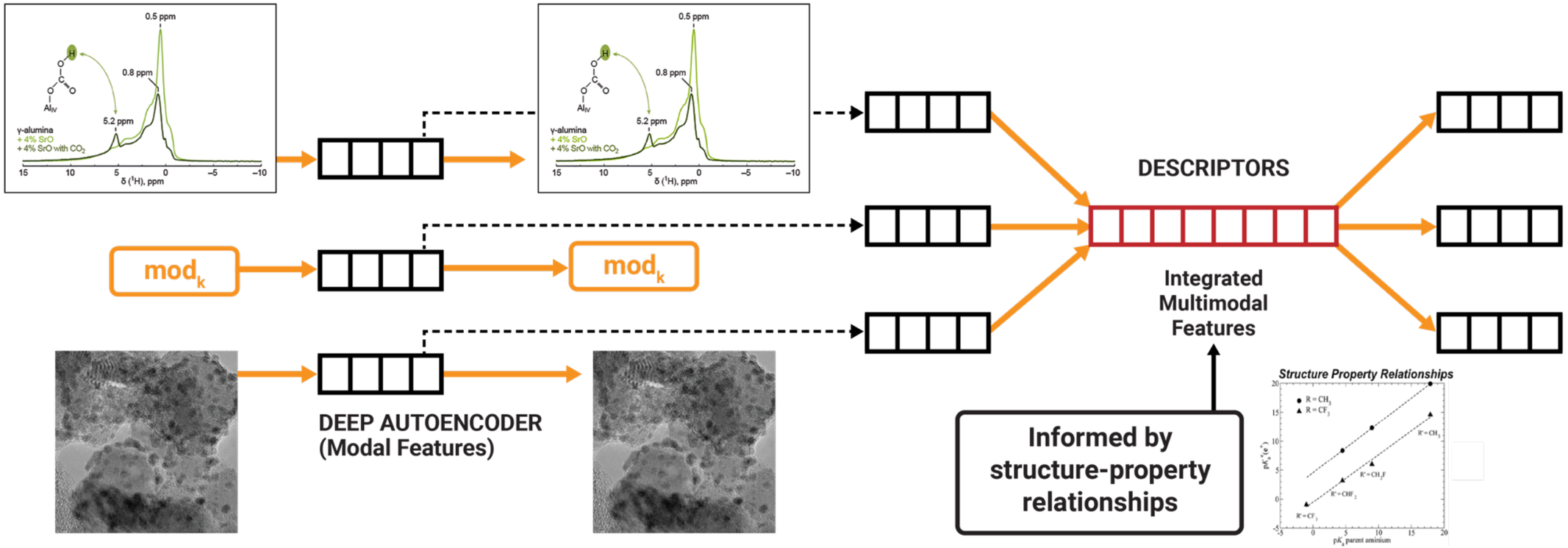


t-sne projection: (a) raw data; (b) autoencoder; (c) clustering ; (d) classification

H. Hu (UW), Y. Liu (UW), Y. Chen (UW), R. Rallo (PNNL)

Integrated multi-modal features

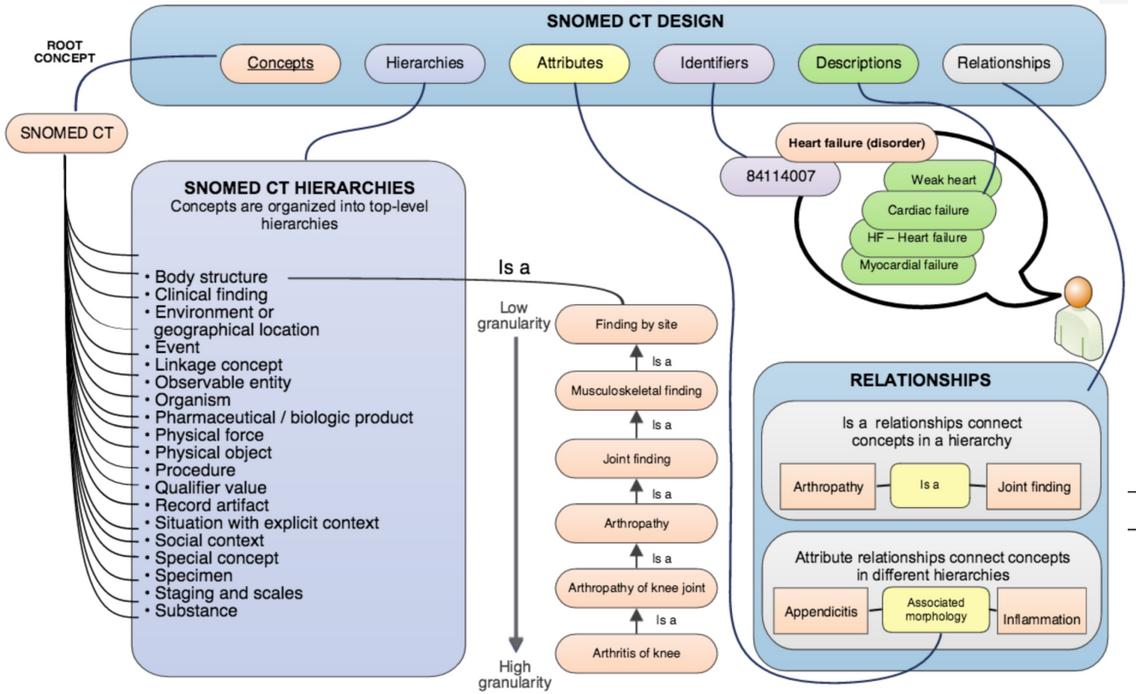
- Heterogeneous data
- Domain-informed stacked autoencoders → integrated features

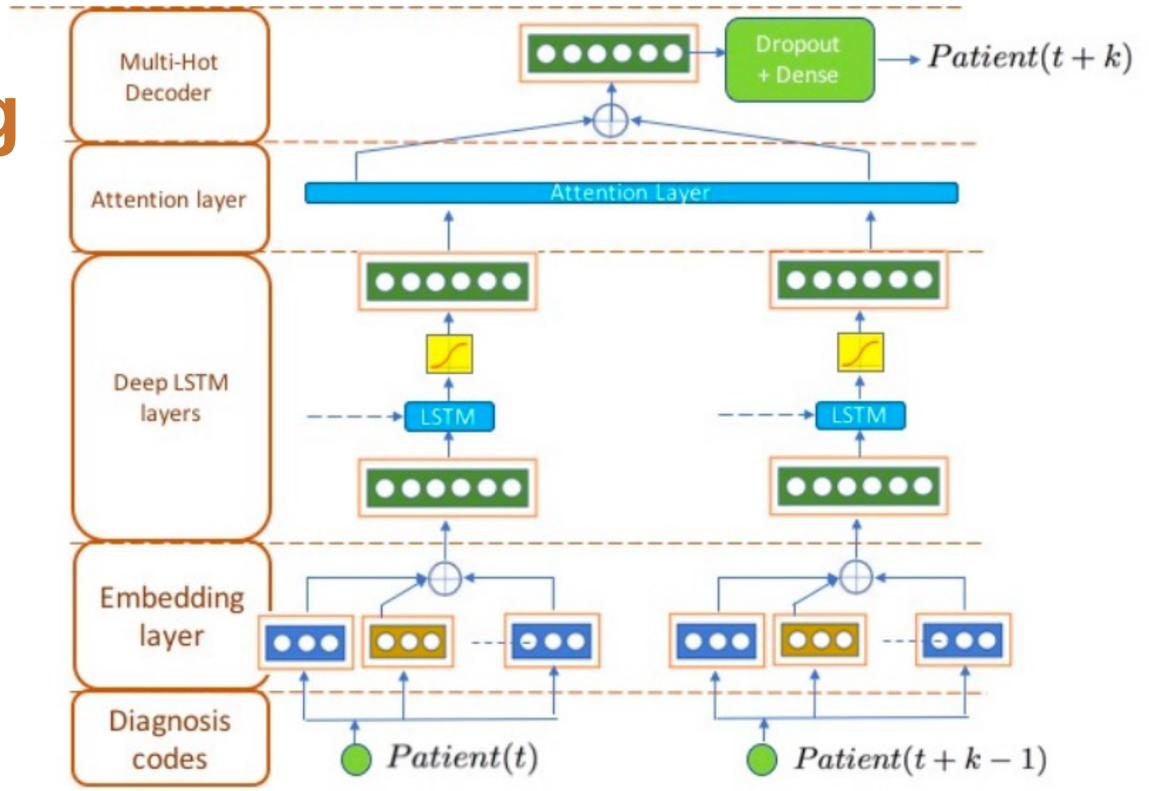


$$L = L_r + L_{SPR}$$

Knowledge embedding

- Medical knowledge / Knowledge graphs





	Node2vec	Metapath2vec	Poincare	CUI2vec	Med2vec
Node Classification	0.817	0.3287	**0.8579**	0.5685	0.0409
Link Prediction	**0.986**	0.3988	0.7135	0.7222	0.8665
Concept Similarity (D1)	**0.79**	0.3	0.7	0.16	NA
Concept Similarity (D3)	**0.90**	0.46	0.31	0.15	NA
Concept Similarity (D5)	**0.81**	-0.32	-0.06	-0.01	NA
Patient State Prediction (All Diagnosis)	0.3938	0.3359	**0.4197**	0.3948	0.3881
Patient State Prediction (Frequent 20)	0.8465	**0.9749**	0.85	0.8035	0.7980
Patient State Prediction (Rare 20)	0.018	0.001	**0.019**	**0.019**	0.011

K. Agarwal, T. Eftimov, R. Addanki, S. Choudhury, S. Tamang, R. Rallo, KDD – Workshop DSHealth, 2019 (https://arxiv.org/abs/1907.08650)

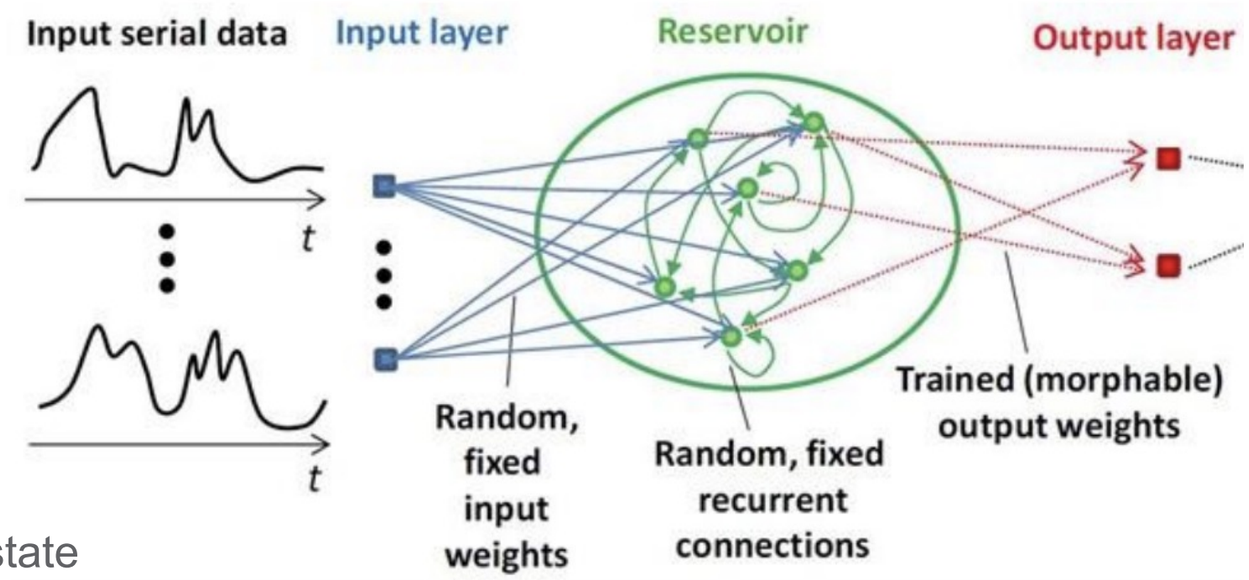
Model structure

- Reservoir computing
  - Generalization of recurrent neural networks
  - Dynamical systems
  - Maps inputs onto a high-dimensional space
  - Hardware implementation
- Elements
  - Input layer
    - ✓ random weights
  - Reservoir
    - ✓ Random sparse connectivity
    - ✓ Non-linear activation
  - Readout layer
    - ✓ Linear transformation of the reservoir state
    - ✓ Fast adaptation using ridge regression

Domain-awareness

Structured reservoir

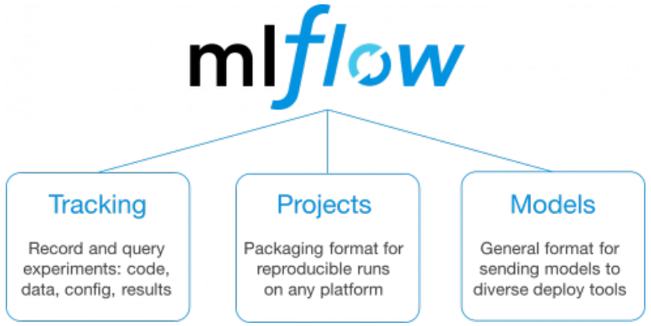
Graph analytics
Network Science

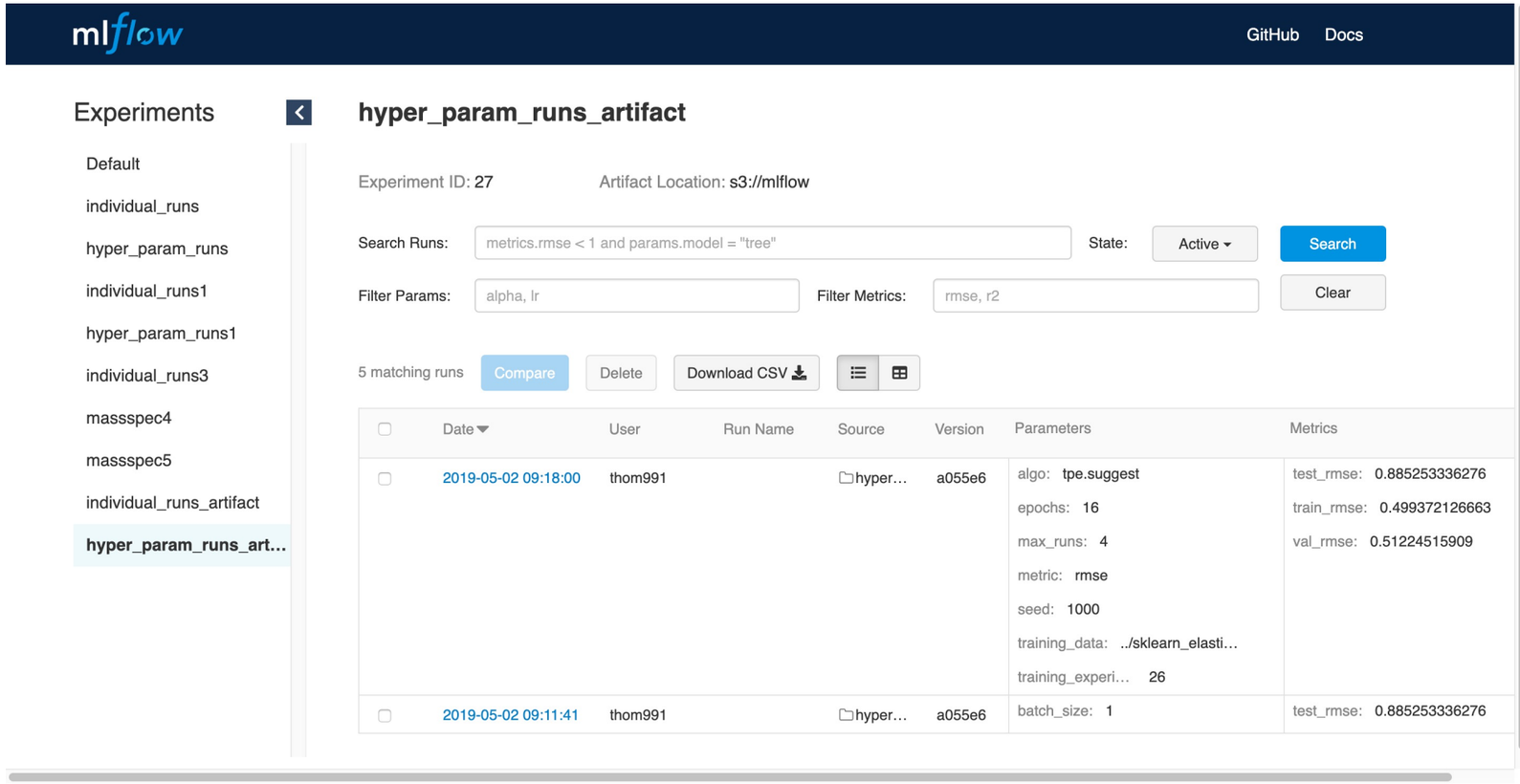


Reproducibility

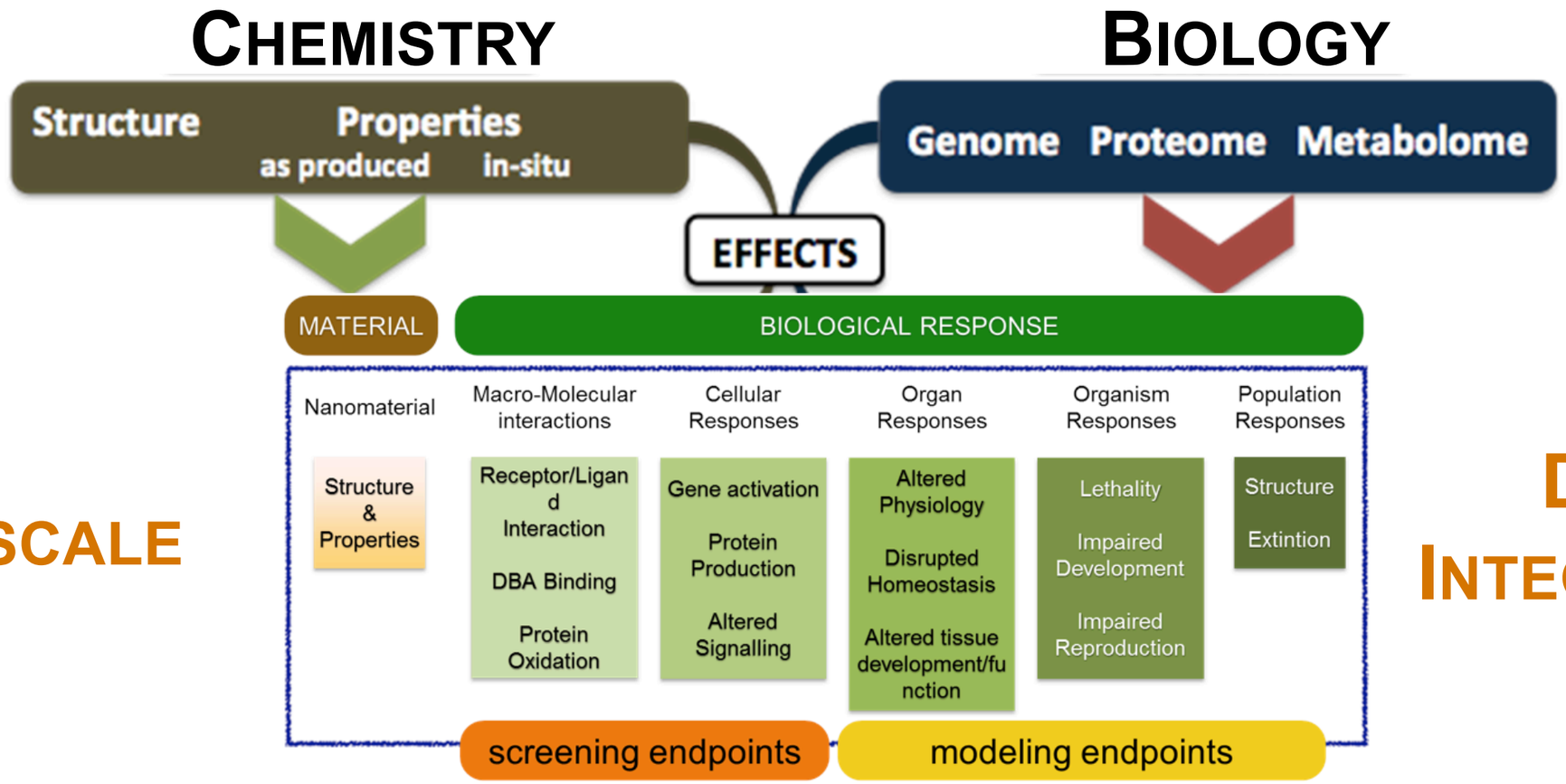
- Provenance tracking in AI/ML workflows

- Metadata on:
  - Task
  - Data
  - Algorithms
  - (Hyper)parameters
  - Performance
- Artifacts
- Packaging
- Deployment





Interpretability: Integrated Domain-informed frameworks



MULTISCALE

DATA INTEGRATION

Multiple levels of biological organization

Roca CP, Rallo R, Fernandez A, Giralt F. Nanoinformatics for Safe-by-Design Engineered Nanomaterials. In *Towards Efficient Designing of Safe Nanomaterials: Innovative Merge of Computational Approaches and Experimental Techniques*, Leszczynski J and Puzyn T (Eds.), 2012, RSC Publishing

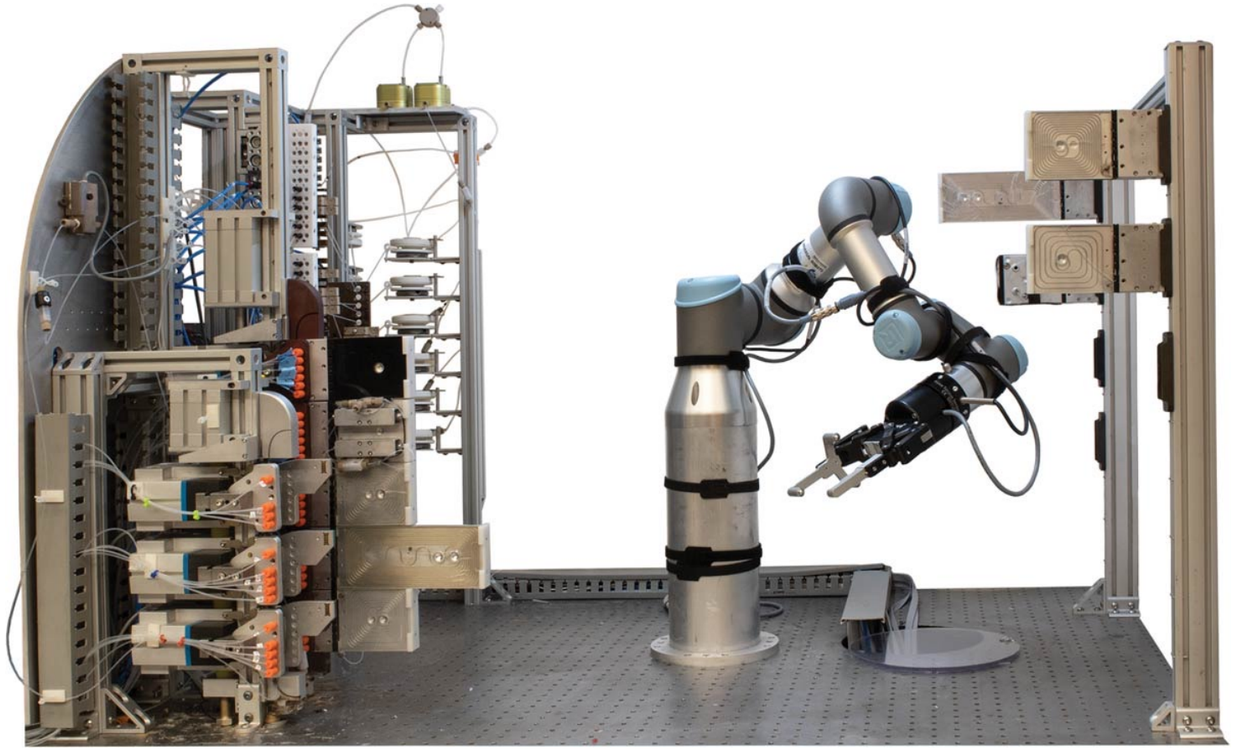
Automation

Interfacing with scientific instruments

- Edge computing for in-situ processing.
- Opportunity to leverage existing real-time control and optimization capabilities.
- Optimal exploration of large combinatorial spaces.

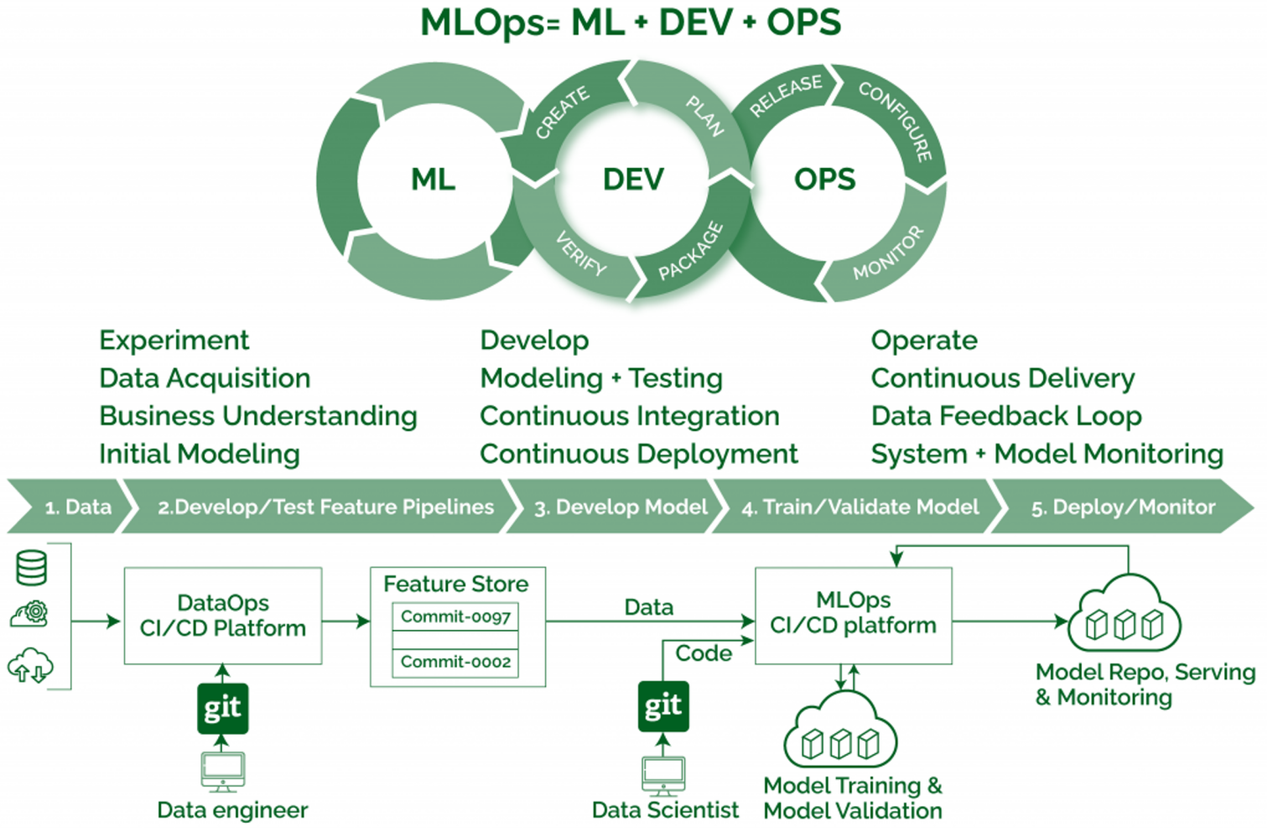
Beyond automation: autonomy

- Hypothesis generation, validation and refinement.
- From machine learning to machine reasoning



Scalability and Deployment

- Rethinking software engineering
  - Composable software systems
  - Heterogeneous computing
- Connecting “Software 1.0” with “Software 2.0”
- Managed and scalable deployment of AI/ML models
  - Lifecycle management
  - Reproducible results
  - Model governance
  - Monitoring and operational use



Adapted from:
[The fundaments of MLOps – The enabler of quality outcomes in production environments]

Summary

Better understanding of bio-nano interactions

Enhanced predictive capabilities

new hypothesis

DISTRIBUTED HIGH-THROUGHPUT COMPUTING

DATA GENERATION
- Synthesis
- Physicochemical characterization

Raw data

DATA MANAGEMENT
- Data Repositories
- Data annotation
- Data quality
- Ontologies
- Provenance
- Data governance

Curated data

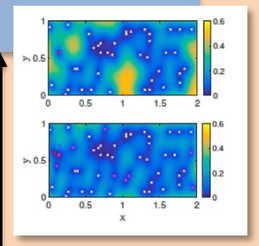
KNOWLEDGE EXTRACTION
- Data mining
- Hypothesis generation

new hypothesis

MATERIALS MODELS

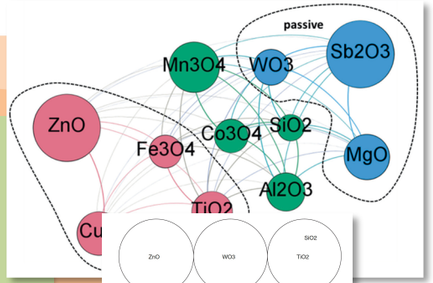
Data-driven
- Machine Learning
- Deep Learning

Simulation
- Multiscale/multiphys
- Multifidelity



Active learning

Optimal design of experiments



SCIENTIFIC COLLABORATION PLATFORM

Pacific Northwest
NATIONAL LABORATORY

Thank you