

Interpretability in AI/ML for Chemical Sciences

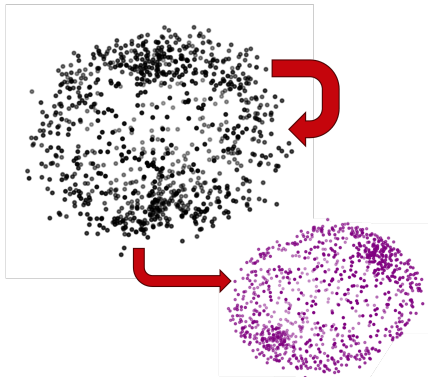
Rose K. Cersonsky

Outline of today

- From representations to predictions
- Feature-based explanations
- Model-based interpretations
- Mapping and chemical space visualization
- Structure–property mapping

Predictive models in chemical informatics focus on connecting materials chemistry, structure, or processing to some set of target quantities.

- At its core, machine learning is about using mathematical operations to identify relationships in a data space.
- We can use these relationships to predict targets, make suggestions of new data points that optimize some value, or understand the space of the data itself.



Write “unsupervised learning” (relationships between data and itself) and “supervised learning” (relationships between data and a target)

Transforming chemistry into numerical formats often abstracts chemical information into a data-rich format.

Data → Representation → Model → Property

Category	Examples / Description
Chemistry-based	Interpreting chemical identity via embeddings or descriptors; Molecular graphs encoding connectivity; One-hot / presence–absence of functional groups
Interaction-based	Based on atomic or molecular positions; Captures many-body environments; Sensitive to conformation and geometry
Implicit (Latent)	Learned representations from neural networks; Encoded in latent space of models; Can be generative or invertible

What is Interpretability?

- **Interpretability:** mapping from mathematical model behavior back to scientific meaning.

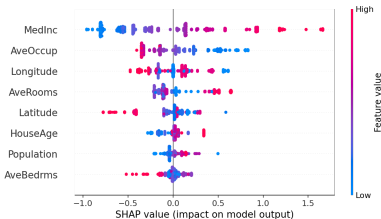
Distinguish: - Explanation (engineering) - Insight (science)

SHAP Values

Shapley value: is a method (solution concept) for fairly distributing the total gains or costs among a group of players who have collaborated.

$$\phi_j = \sum_{S \subseteq N, j \notin S} \frac{|S|!(|N| - |S| - 1)!}{|N|!} (V(S \cup \{j\}) - V(S))$$

- ϕ_j : contribution of player j
- S : a coalition containing j
- N : all players
- $(V(S \cup \{j\}) - V(S))$: the marginal contribution of player j to coalition S
- $\frac{|S|!(|N| - |S| - 1)!}{|N|!}$: weight of the marginal contribution



The central idea is to fairly distribute the prediction among the input features.

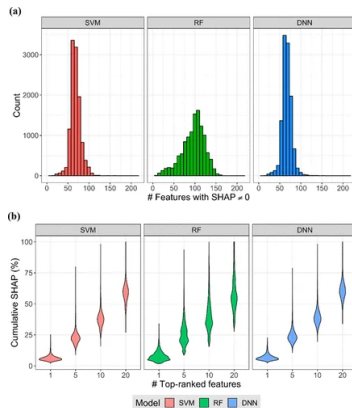
In this framework, each feature is treated as a “player,” and the model output is treated as the “payout” of a game.

The goal is to determine how much each feature contributes to that payout.

Now, let's interpret this more concretely.

This paper introduces SHAP (Shapley Additive Explanations) as a model-agnostic method to interpret machine learning predictions of compound bioactivity, showing how complex models such as random forests, SVMs, and deep neural networks can be decomposed into feature-level contributions that can be mapped back onto molecular structures. By applying SHAP to structure–activity relationship (SAR) models, the authors demonstrate that important substructures driving activity predictions can be identified consistently across different models, enabling chemists to rationalize predictions, validate models against domain knowledge, and potentially guide compound design. However, the work also highlights that these explanations remain tied to the chosen feature representation, meaning interpretability is limited by how chemical information is encoded, even if the attribution itself is mathematically rigorous.

J. Med. Chem. 2020, 63, 16, 8761–8777

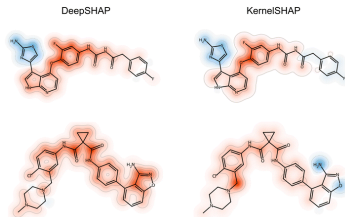


Limitations of Shapley Analysis

they attribute importance at the level of input features rather than underlying scientific variables, meaning the explanation is only as meaningful as the chosen representation;

when features are correlated, redundant, or abstract (e.g., fingerprint bits or latent variables), the computed contributions can be unstable or misleading because Shapley values rely on averaging marginal effects across hypothetical feature subsets that may not correspond to physically realizable systems.

In addition, the explanations are local—specific to individual predictions—and do not inherently capture global structure–property relationships, while also being computationally expensive and dependent on assumptions (such as feature independence or background distributions) that often do not hold in chemical data, limiting their ability to yield direct mechanistic insight.



Chem. Sci., 2026, 17, 1411-1422

Saliency Maps

Saliency map: highlights either the region on which people's eyes focus first or the most relevant regions for machine learning models

$$S(x) = \frac{\partial f(x)}{\partial x}$$



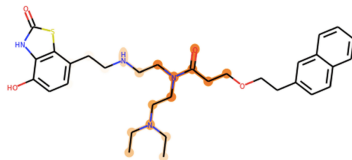
Saliency maps use gradients to measure sensitivity of predictions to input changes. They are widely used in neural networks.

Now, let's interpret this more concretely.

explainable graph convolutional neural network (gCNN) architecture for small molecule activity prediction.

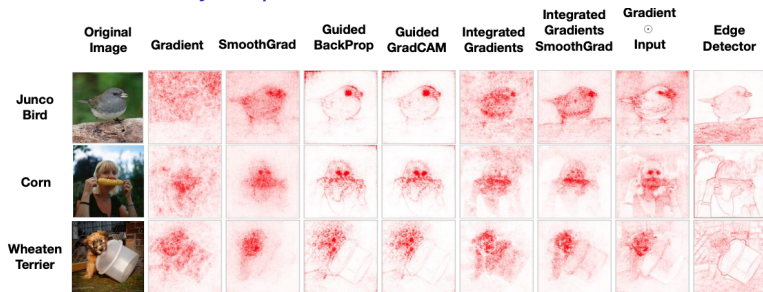
800 protein targets that produces a simplified gCNN QSAR architecture, and we observe that such a model can yield performance improvements over both standard gCNN and RF methods on difficult-to-classify test sets.

saliency map technology that highlights molecular substructures relevant to activity, and we perform saliency analysis on nearly 100 data-rich protein targets.



J Comput Aided Mol Des 36, 391–404 (2022).

Limitations of Saliency Maps



saliency maps provide only local, infinitesimal sensitivity—they quantify how small perturbations affect a prediction, but chemical changes are typically discrete (e.g., functional group substitutions), meaning gradients do not correspond to realistic chemical transformations.

saliency is highly sensitive to input scaling, model architecture, and noise, which can produce visually striking but inconsistent importance maps.

gradients are taken with respect to atom or bond embeddings, which are often abstract latent quantities, so mapping them back to interpretable structural motifs requires additional heuristics, limiting their ability to yield robust, mechanistic insight about structure–property relationships.

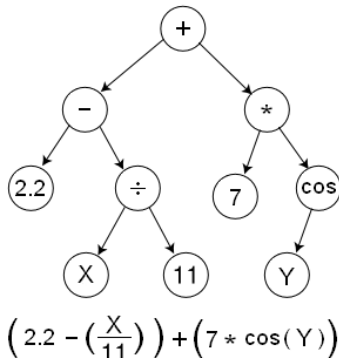
expensive!

Proceedings of the 32nd International Conference on Neural Information Processing Systems
Pages 9525 - 9536

Symbolic Regression

Symbolic Regression: searches over mathematical expressions to discover an explicit functional relationship between inputs and outputs.

$$f(x) = x_1^2 + \log(x_2) - \frac{x_3}{x_4}$$



Symbolic regression takes a fundamentally different approach to interpretability.

The goal is to find an equation that both fits the data well and remains relatively simple.

This is typically achieved using evolutionary or genetic programming algorithms.

Candidate equations are generated, evaluated, and iteratively improved.

The key advantage is that the output is an explicit equation.

This means the model itself is interpretable, rather than requiring a post hoc explanation.

Limitations of Symbolic Regression

Despite its interpretability advantages, symbolic regression has several important limitations.

First, the search space of possible equations is extremely large, which makes the optimization problem computationally challenging.

Second, there is often a tradeoff between accuracy and simplicity.

Highly accurate models may become algebraically complex, reducing interpretability.

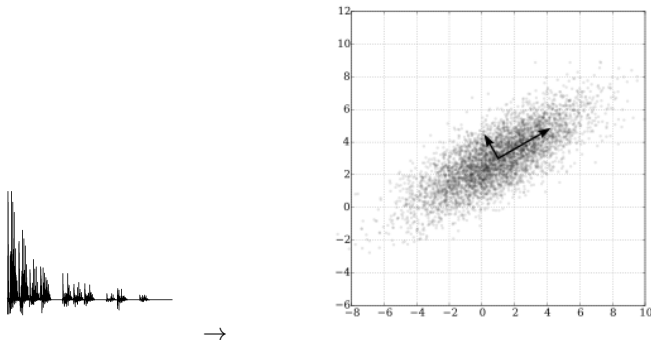
Third, the discovered equations depend strongly on the chosen input features.

If the feature set does not contain physically meaningful variables, the resulting expression may still be difficult to interpret.

Fourth, symbolic regression can be sensitive to noise in the data, which may lead to unstable or non-robust expressions.

So while it is powerful for discovering explicit relationships, it is best used in combination with other interpretability methods.

“Seeing” data typically corresponds to unsupervised mapping: Mapping is a subset of machine-learning methods that focus on visualizing high-dimensional data in human-interpretable dimensions.

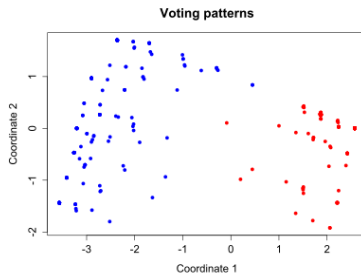
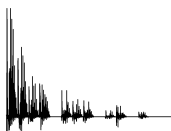


Principal Components Analysis: linear projection that finds orthogonal directions

$$T = XW, \quad W = [v_1, \dots, v_k]$$

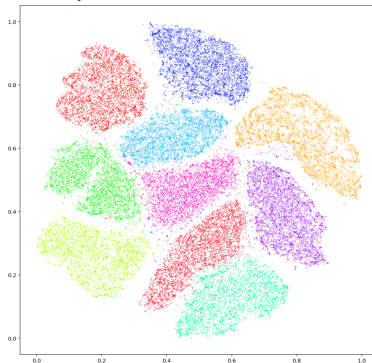
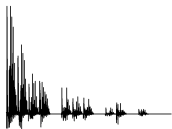
where v_i are the eigenvectors of the covariance

“Seeing” data typically corresponds to unsupervised mapping: Mapping is a subset of machine-learning methods that focus on visualizing high-dimensional data in human-interpretable dimensions.



Multidimensional Scaling: embedding method that finds points $\mathbf{T}$ that preserve pairwise distances $\|\mathbf{t}_i - \mathbf{t}_j\| \approx d_{ij}$

“Seeing” data typically corresponds to unsupervised mapping: Mapping is a subset of machine-learning methods that focus on visualizing high-dimensional data in human-interpretable dimensions.

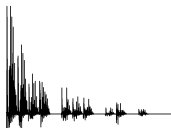


t-Stochastic Neighborhood Embeddings
(t-SNE): nonlinear method that matches

$$T = \arg \min_Z \text{KL}(P_{ij} \parallel Q_{ij})$$

where P_{ij} and Q_{ij} are high- and low-dimensional similarity distributions.

“Seeing” data typically corresponds to unsupervised mapping: Mapping is a subset of machine-learning methods that focus on visualizing high-dimensional data in human-interpretable dimensions.

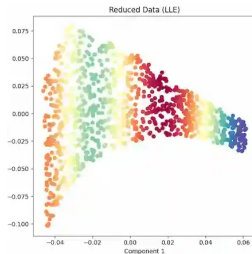
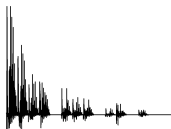


Uniform Manifold Approximation and Project (UMAP): graph-based manifold embedding that optimizes cross-entropy between fuzzy simplicial sets

$$T = \arg \min_Z \sum_{i,j} [p_{ij} \log q_{ij} + (1 - p_{ij}) \log(1 - q_{ij})]$$

optimizes a cross-entropy between high- and low-dimensional graph similarities.

“Seeing” data typically corresponds to unsupervised mapping: Mapping is a subset of machine-learning methods that focus on visualizing high-dimensional data in human-interpretable dimensions.

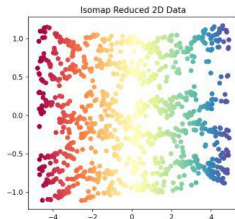
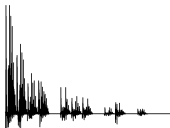


Locally-linear Embedding: manifold method that preserves local linear reconstructions

$$T = \arg \min_Z \sum_i \left\| z_i - \sum_j w_{ij} z_j \right\|^2$$

preserves local linear reconstruction weights w_{ij} .

“Seeing” data typically corresponds to unsupervised mapping: Mapping is a subset of machine-learning methods that focus on visualizing high-dimensional data in human-interpretable dimensions.



Isomap: nonlinear dimensionality reduction that preserves geodesic (graph-based) distances

$$T = \arg \min_Z \sum_{i,j} \left(\|z_i - z_j\| - d_{ij}^{\text{geo}} \right)^2$$

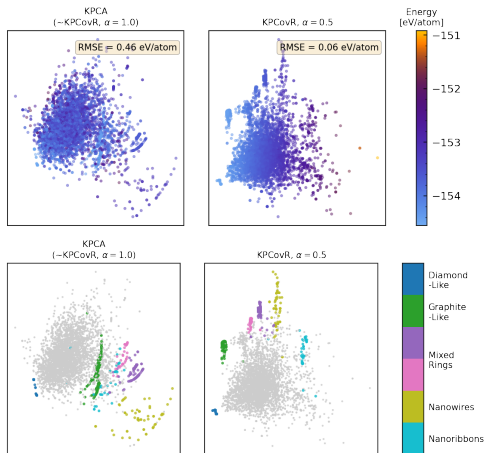
embeds data to preserve geodesic distances d_{ij}^{geo} on a neighborhood graph.

Multi-objective Structure–Property Mapping

$$\ell = \alpha \ell(T \rightarrow X) + (1 - \alpha) \ell(T \rightarrow Y)$$

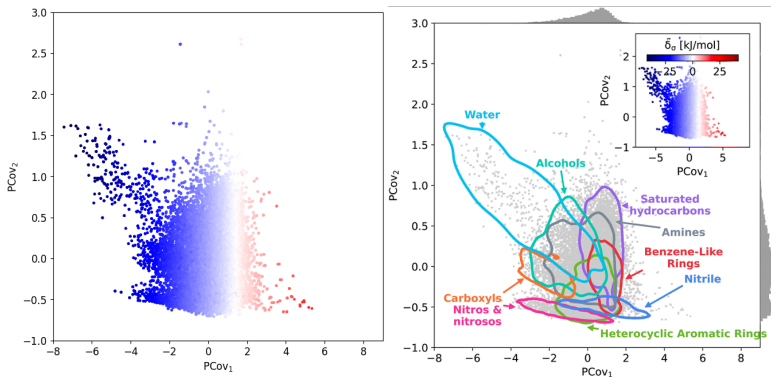
- Mapping chemical data involves competing objectives
- Preserve structural diversity (geometry of X)
- Capture correlations with target properties (Y)
- α controls tradeoff between **structure** and **property**
- Embedding T reflects both similarity and prediction relevance

Structure vs Property



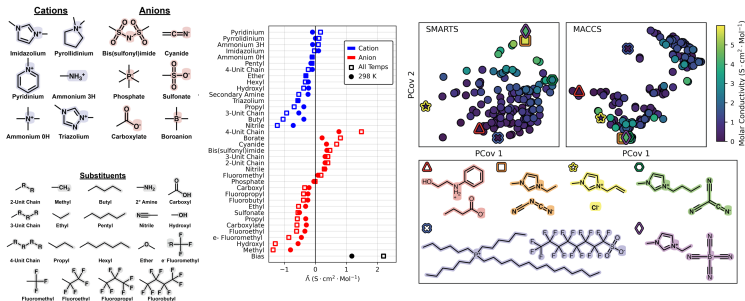
- Unsupervised embeddings (e.g. PCA, KPCA) capture variation but not relevance
- Structure may dominate signal even when property differs
- PCovR introduces supervision into the embedding
- Produces maps aligned with **scientifically meaningful variation**

Functional Group Contributions to Crystallization



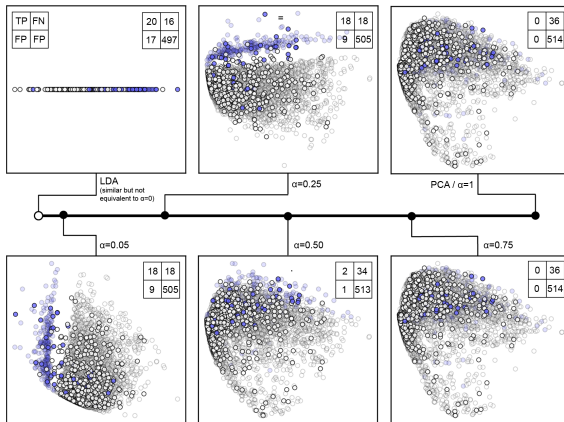
- Each point corresponds to a molecular fragment or motif
- Color encodes contribution to stability
- Blue: stabilizing, Red: destabilizing
- Enables ranking of functional groups by effect
- Direct link between representation and chemical insight

Connecting Structure to Emergent Behavior in Ionic Liquids



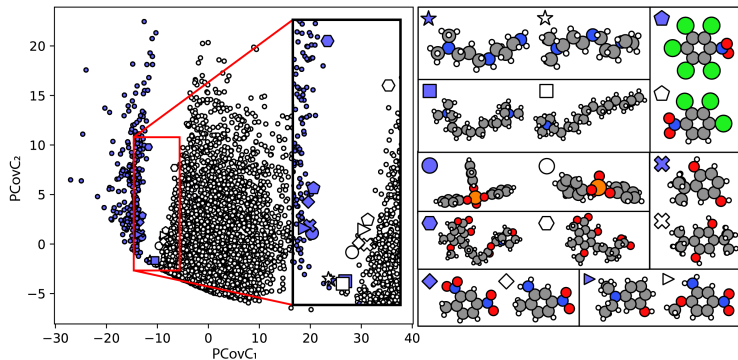
- Embeddings capture links between microscopic structure and macroscopic properties
- Enables regression and classification tasks simultaneously
- Predictions validated via observed correlations
- Supports discovery of emergent chemical trends

Visualizing Decision Boundaries

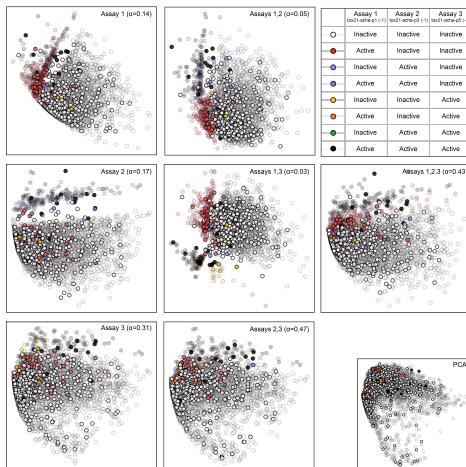


- Latent space enables visualization of classifier behavior
- Boundaries separate predicted classes
- Geometry reveals margin, overlap, and uncertainty
- Useful for debugging models and understanding errors

A little bit more in depth



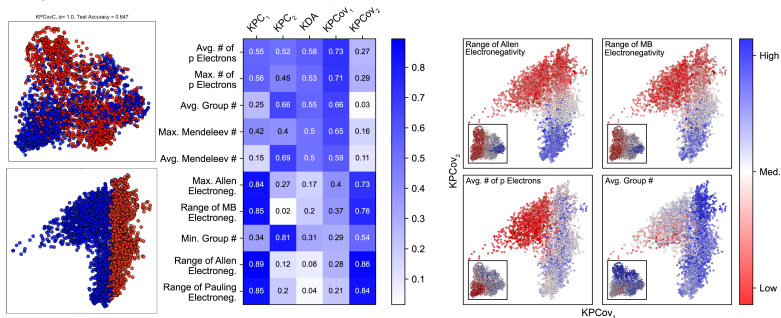
Handling Conflicting Labels



- Real datasets often contain inconsistent labels
- Multiple assays or measurements may disagree
- Embedding reveals regions of disagreement
- Enables identification of ambiguous or noisy data

Nonlinear Relationships in Chemistry

Dataset: 4,916 inorganic compounds
Feature Set: 136 elemental properties
Target: Metal (blue), nonmetal (red)



Key Takeaways

- Interpretability operates at multiple levels
- Feature attribution explains *model decisions*
- Mapping methods reveal *structure-property relationships*
- Symbolic regression discovers *explicit equations*

To conclude, interpretability in chemical machine learning is not a single concept— it operates at multiple levels.

Feature attribution methods, like SHAP and saliency, focus on explaining individual predictions.

They tell us what the model is doing, but not necessarily what the chemistry is doing.

Mapping methods, such as PCA and PCovR, shift the focus to the structure of data itself.

These methods allow us to visualize relationships in chemical space, and understand how properties emerge from structure.

Symbolic regression takes yet another perspective, producing explicit mathematical equations that describe relationships in the data.

Each of these approaches provides a different lens.

In practice, true scientific insight often comes from combining these approaches.

The goal is not just to build accurate models, but to use them as tools for discovery.