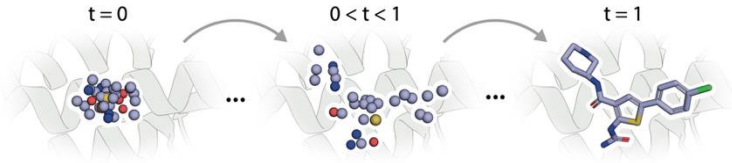
# GNNs for Drug Discovery

Adrian Dobbelstein
Tian Zhu
LPDI


10.12.2025

# Own research: Generative Modelling for Structure-Based Drug Discovery




High-throughput screening library

Virtual screening library

Generative modelling?

Organic compounds < 0.5kDa

≈$10^6$ ≈$10^8$ ≈$10^{60}$

Size of covered chemical space

Visualization :10.12688/f1000research.21540.2

# LDDM: Large Drug Discovery Model

## LLM

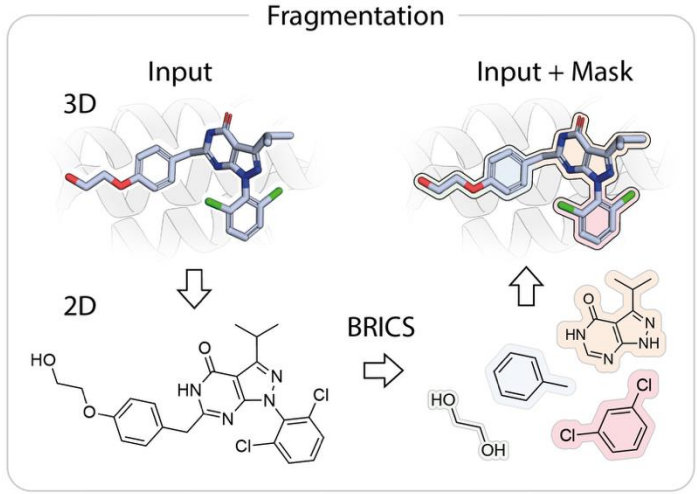
Lausanne is a city near Lake Geneva

Lausanne is a [MASK] near Lake Geneva

↓ Predict

Lausanne is a city near Lake Geneva

## LDDM

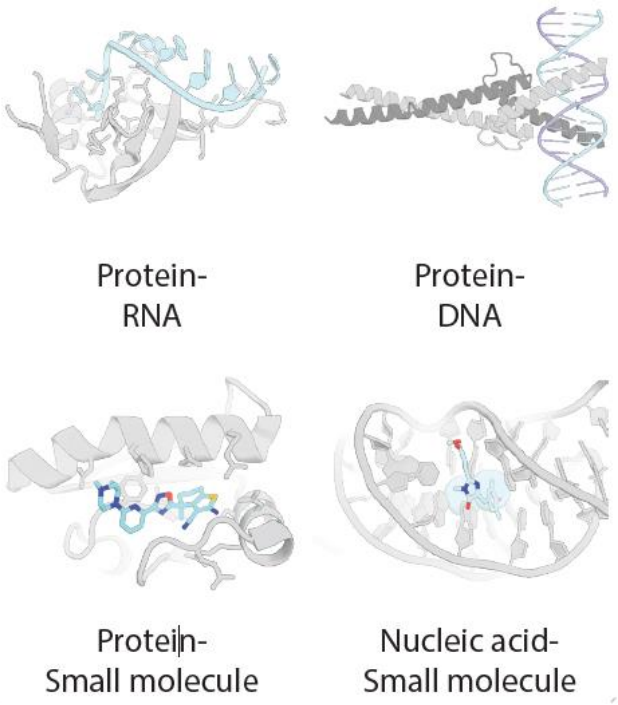



# Learning Universal Representations of Intermolecular Interactions with ATOMICA

Fang, A., Desgagné, M., Zhang, Z., Zhou, A., Loscalzo, J., Pentelute, B.L. and Zitnik, M., 2025.

# Can we learn a universal representation of biomolecular interactions?

- ML for Drug Discovery: Most models focus on a *single* modality: Small molecules (SM), Proteins, Peptides, DNA/RNA
- Differences across modalities:
  - Atom composition
  - Molecular properties
  - Training data abuncance
- Interaction types found in all modalities, the composition however is modality-specific:
  - Protein-protein interactions (PPIs): “Global” geometric, hydrophobic, and electrostatic complementarity; large surface area
  - Protein-SM interactions: Highly specific; smaller surface area
  - Metal ions: Coordination geometry



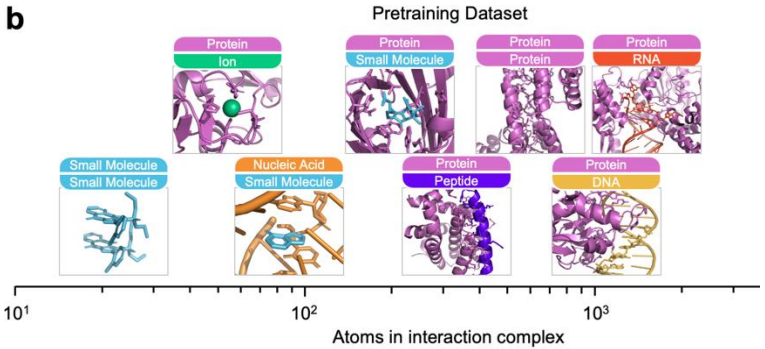

ATOMICA: Can we learn representations of biomolecular interactions *across modalities?*

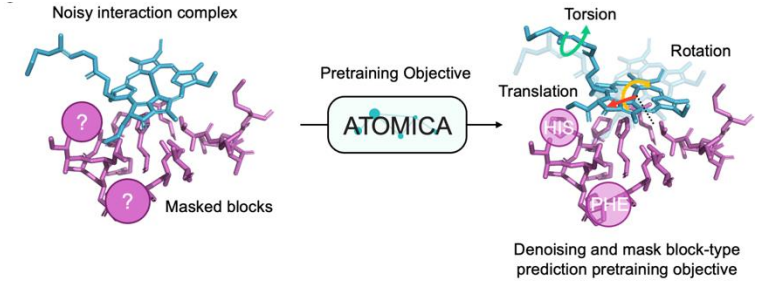
# Overview

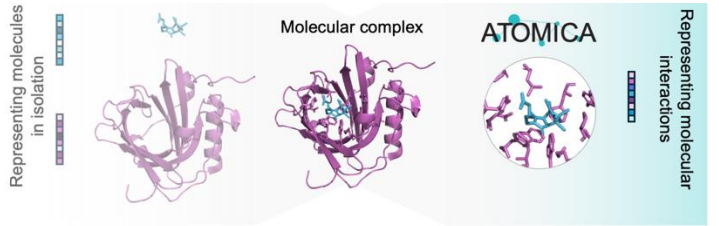
- Diverse dataset of diverse modalities

SM-SM, Protein-*

- Complexes are tokenized and processed with a hierarchical MPNN
- Self-supervised pretraining:
  - Predict masked block (token) identity.
  - Reconstruct geometric pertubations.
- Block- and Atom-level feature vectors are useful for various downstream tasks.

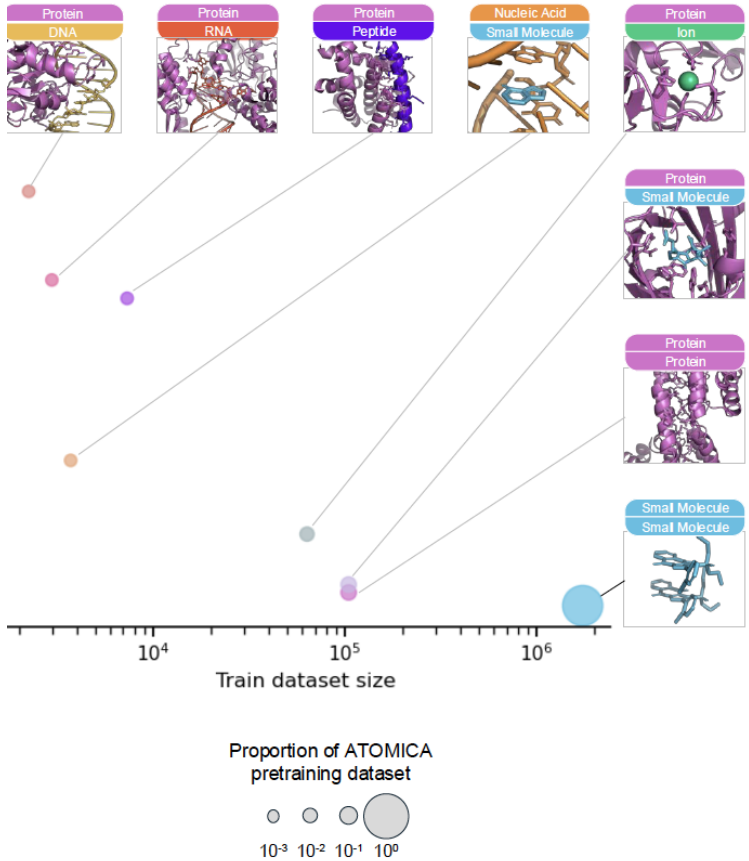


# Training data

- Small molecule – Small molecule interactions:
  - Crystal contacts from Cambridge Structural Database
  - ~1.7 x $10^6$ datapoints
  - Fingerprint-based split
- Q-BioLiP: Derived from PDB, protein/peptide/nucleotide/small molecule/ions
  - 3.4 $10^5$ datapoints
  - Sequence identity-based split



# Data representation

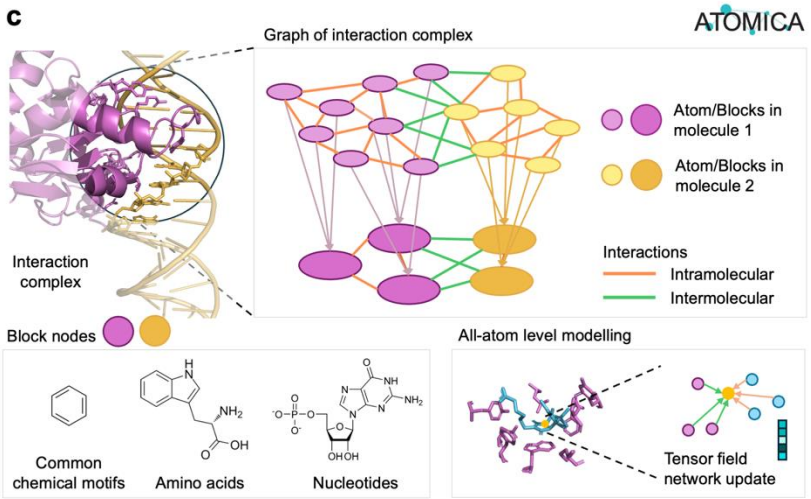
- Data represented as hierarchical graph:
  1. Atoms ≙ Nodes
  2. Residues/Nucleotides/Fragments ≙ Nodels

Each atom/block has an identity + coordinates:

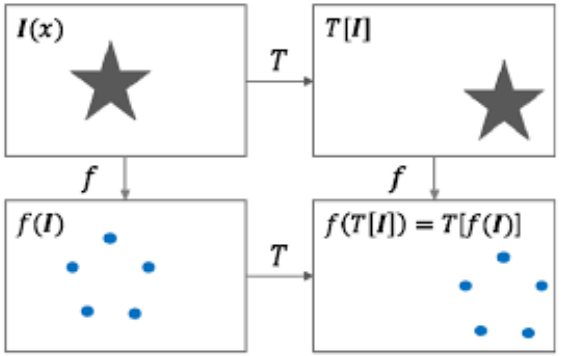
- 20 amino acid block, 4+4 DNA/RNA nucleotides, 290 small molecule fragments, 118 singleton-atom block types
- Special blocks: [Mask], “Unkown”
- Virtual nodes: “Global” block connected to all other blocks

Graph construction:

- Atom nodes communicate with parent block nodes.
- kNN graph for intra- *and* intermolecular edges for *both* atom-level and block-level graph



# GNN architecture




- SE(3)-equivariant MPNN
- Representation learning for atom-level, block-level, and graph-level.
- Atom level: Uses DiffDock-based MPNN architecture.
- Atom types/Block types: Embedded into continous feature vectors.
- Message passing:
  - Sperical embedding of relative atom positioning: $Y(r_{ab})$.
  - Scaling factor $\psi_{ab}$ dependent on inter- and intramolecular edge type ($t_{ab}$).
- Final MLP after MP layers to obtain $h^{atom}$.

$$\mathbf{h}_a^{\text{atom}} \leftarrow \mathbf{h}_a^{\text{atom}} + \text{LN}\left(\frac{1}{|\mathcal{N}_a|}\sum_{b\in\mathcal{N}_a} Y\left(\hat{\mathbf{r}}_{ab}\right) \otimes_{\psi_{ab}} \mathbf{h}_b^{\text{atom}}\right)$$

$$\text{with } \psi_{ab} = \Psi\left(\mathbf{e}_{ab}, \mathbf{t}_{ab}, \mathbf{h}_{a,(0e)}^{\text{atom}}, \mathbf{h}_{b,(0e)}^{\text{atom}}\right).$$

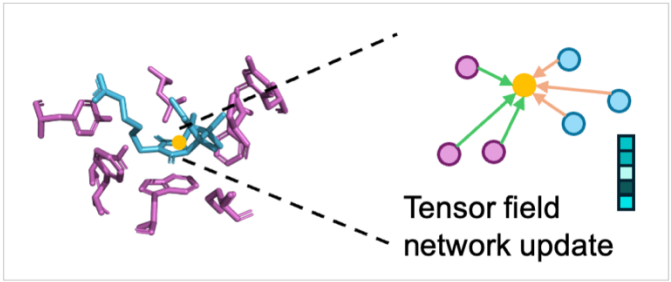
# GNN architecture (2)

- Block-level:
  - Updates *between blocks:* MP as for atoms with separate parameters.
  - Updates from atoms to blocks: Multi-head attention with multiple attention heads.
  - Each attention head aggregates information from the corresponding atom-level features.
- Graph-level: Self-attention on blocks + sum to obtain $h^{graph}$.

$$\mathbf{h}_b^{\text{block}} \leftarrow \mathbf{h}_b^{\text{block}} + \text{MultiHead}(\mathbf{h}_b^{\text{block}}, \{\mathbf{h}_a^{\text{atom}}\}_{a \in A_b})$$

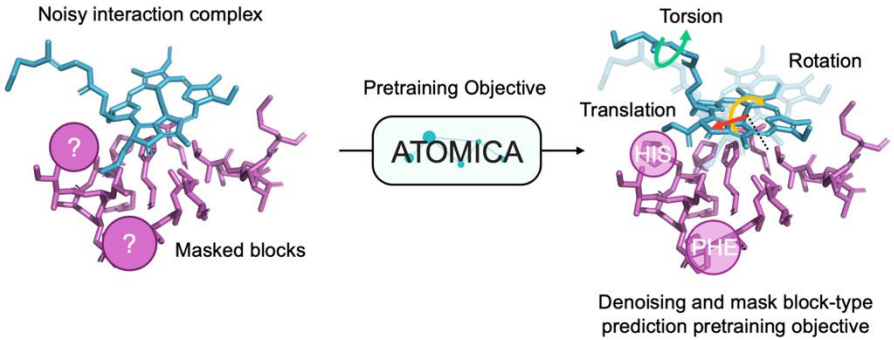
$$\text{MultiHead}(\mathbf{h}_b^{\text{block}}, \{\mathbf{h}_a^{\text{atom}}\}_{a \in A_b}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_{n_{\text{heads}}})\mathbf{W}_O$$

All-atom level modelling



# Training objective: Self-supervised denoising and block identity prediction

- Pertubations:
  - Sampling rotation and translation vectors $\omega \sim N_{SO(3)}$ and $\mathrm{t} \sim \mathrm{N}(0, \sigma_t^2 \mathrm{I})$ for each molecule.
  - Pertubations applied to block and atom coordinates.
  - Torsion angles of m side chains/rotatable bonds are perturbed by $\theta \sim N_{SO(2)^m}$.
- Block identity masking: 10% of block identities assigned the [MASK] block type.




- Denoising perturbations using score-matching:

$$\nabla_{\mathbf{t}} \log p(\mathbf{t}) = -\mathbf{t}/\sigma_t^2$$

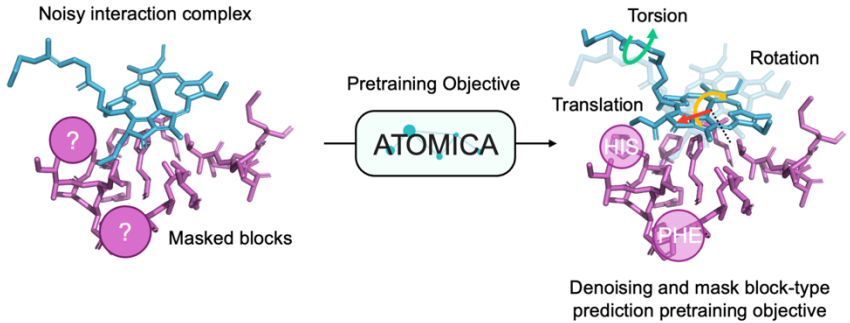
$$l_\omega = ||\mathbf{s}_{\boldsymbol{\omega}} - \nabla_{\boldsymbol{\omega}} \log p(\boldsymbol{\omega})||^2$$

$$l_t = ||\mathbf{s}_{\mathbf{t}} - \nabla_{\mathbf{t}} \log p(\mathbf{t})||^2$$

$$l_\theta = \sum_z ||\mathbf{s}_{\boldsymbol{\theta}_z} - \nabla_{\boldsymbol{\theta}_z} \log p(\boldsymbol{\theta}_z)||^2$$

# Training objective: Self-supervised denoising and block identity prediction (2)

- Prediction of rotation/translation scores uses the representations $h^{atom}$, $h^{block}$, and $h^{graph}$.




- 1-hop MP using the center c of the perturbed molecule A' for rescaling of the tensor product:

$$\mathbf{s} \leftarrow \mathrm{LN}\left(\frac{1}{|\mathcal{A}'|}\sum_{a\in\mathcal{A}'} Y\left(\hat{r}_{ca}\right)\otimes_{\phi_{ca}}\mathbf{h}_a^{\mathrm{atom}}\right) \quad \text{with } \phi_{ca}=\Phi\left(\left\|\mathbf{e}_{ca}, \mathbf{h}_{a,(0e)}^{\mathrm{atom}}\right\|\right)$$

- Final prediction of the rotation/translation score using the global embedding $h^{graph}$:

$$\mathbf{s_t} = \Gamma_{\mathbf{t}}(\mathbf{h}_i^{\mathrm{graph}}) * \mathbf{s}_{(1o)}$$

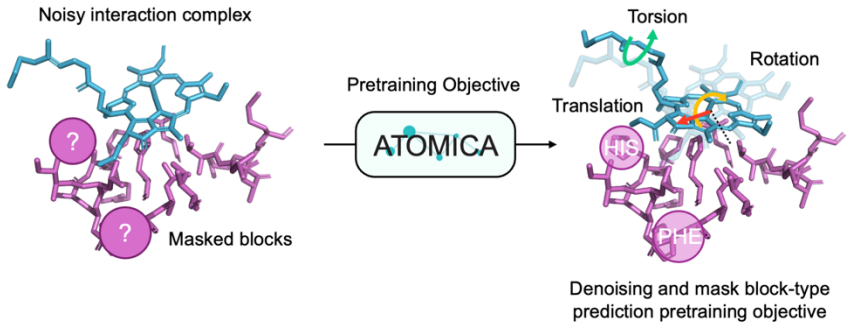
# Training objective: Self-supervised denoising and block identity prediction (2)

- Reconstruction of masking: Predicting probability vectors for each block identity with an MLP on the block embeddings:

$$\hat{\mathbf{y}}_b = \mathrm{Softmax}(\Upsilon(\mathbf{h}_b^{\mathrm{block}}))$$

- Training with CE loss.

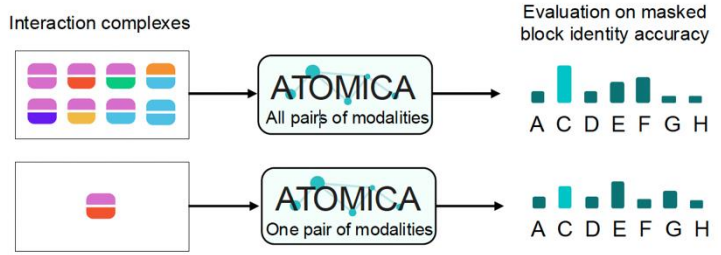
- Total pretraining loss: $\mathcal{L} = \beta_\omega l_\omega + \beta_t l_t + \beta_\theta l_\theta + \beta_m l_m$

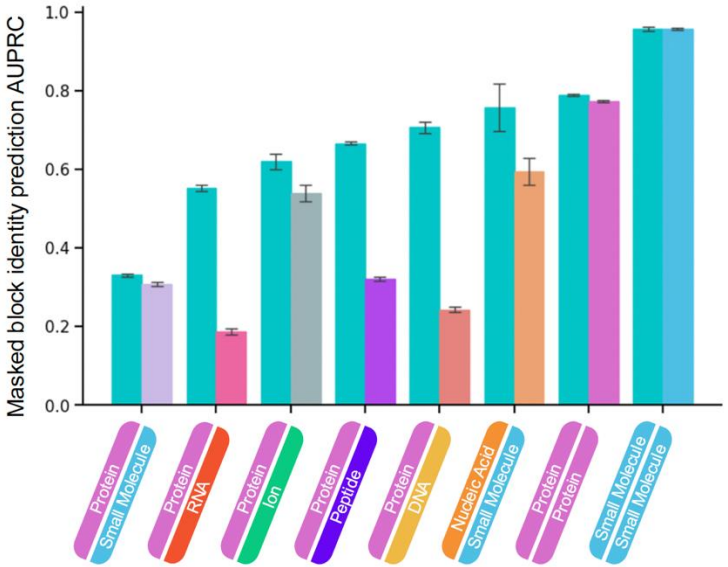


# Multi-modality pretraining leads to performance improvement in rare interaction domains

- Task: Masked block identity prediction in held-out test set.
- Ablation study: Separate models trained on single interaction modalities (e.g., only Protein-DNA).
- Multi-modal pretraining leads to drastic improvement in protein-nucleotide prediction AUPRC.
- ΔAUPRC inversely proportional to the number of datapoints in one modality.

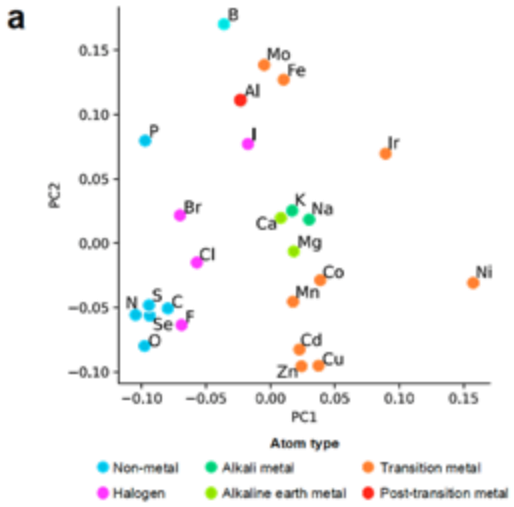
➢ Multimodal pretraining drives generalizability.

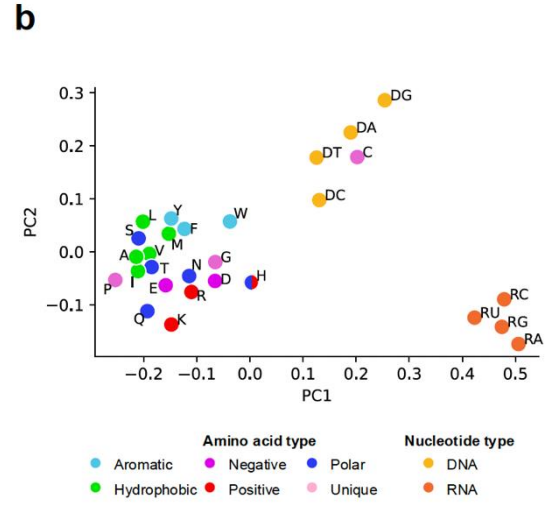


# ATOMICA latent space organizes blocks into biological meaningful units

- Graph-level embeddings separate PPIs from other interaction types.
- Block- and Atom-level embeddings separate atoms resembling their physical properties.

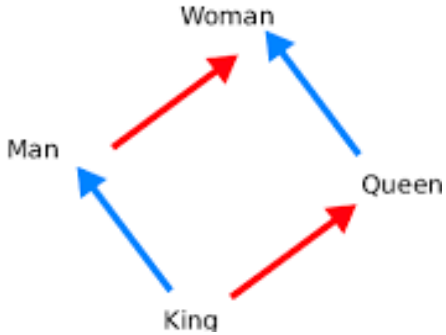


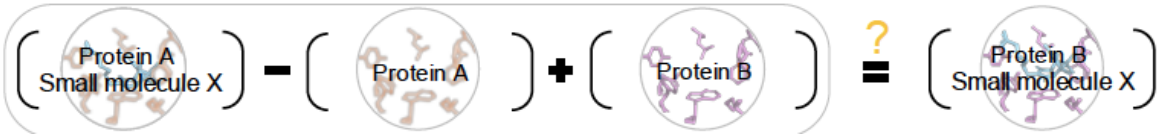
# ATOMICA embeddings follow algebraic relationsships

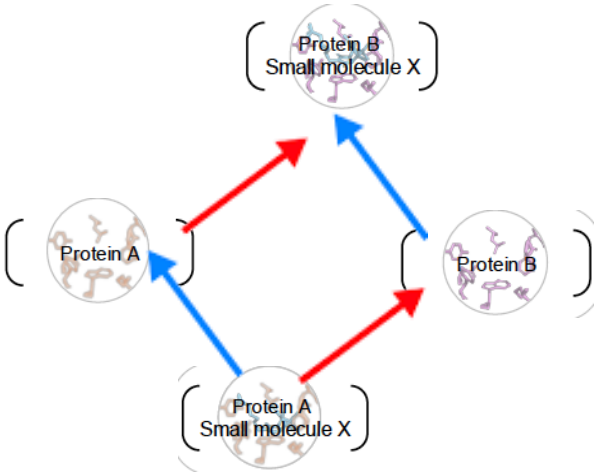
- Word2vec: Word embedding follow algebraic relationships, e.g.,

e(Man) – e(King) + e(Queen) ≈ e(Woman)




- Similar structure for ATOMICA embeddings?



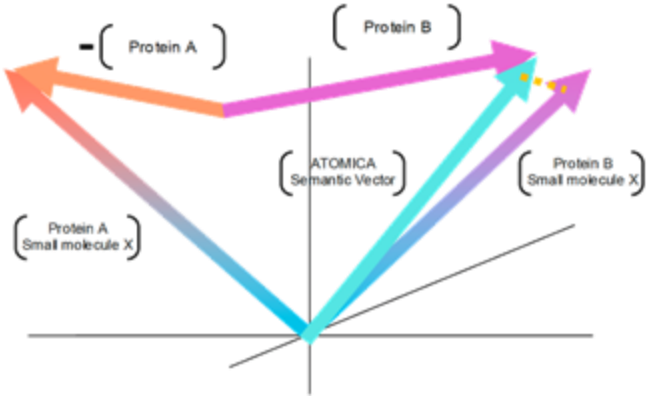





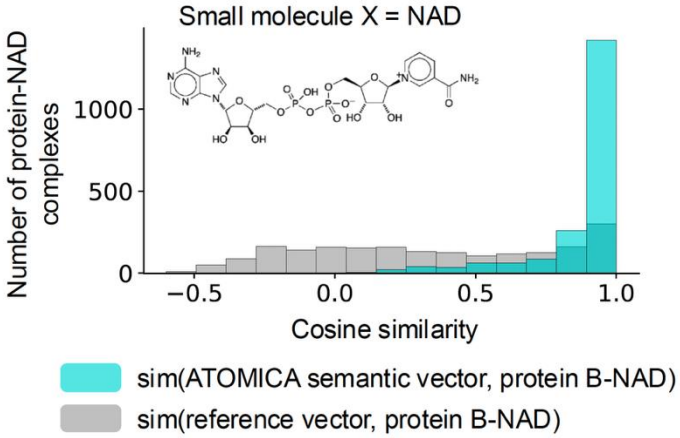

S. Liu *et al. IEEE Transactions on Visualization and Computer Graphics.* 2018

# ATOMICA embeddings follow algebraic relationsships (2)

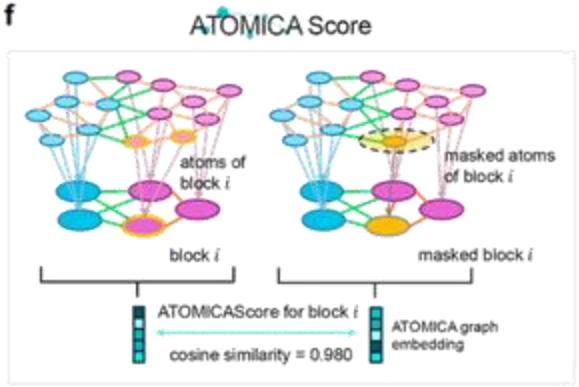
Atomica Semantic vector:

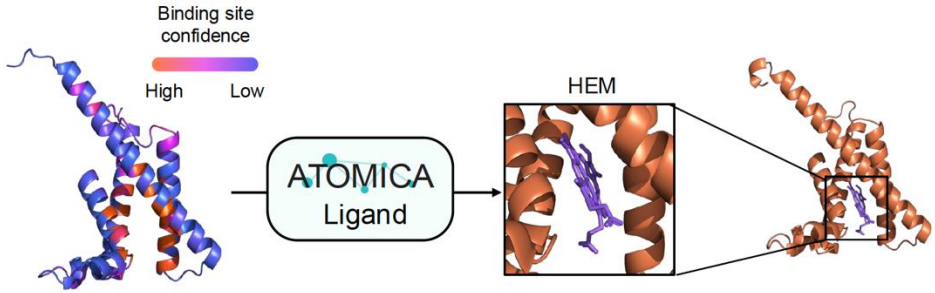

- Known Protein A-Nicotinamide adenine dinucleotide complex embedding.
- Unknown NAD-binding protein B with max. 30% seq similarity: Semantic vectors are close to the true B-NAD embedding.



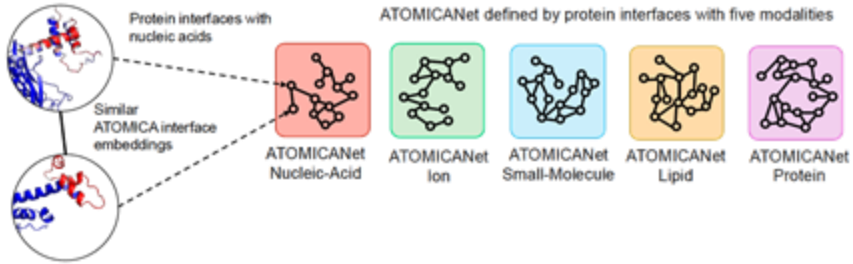
# Overview of applications




Scoring mutation effects in binding sites



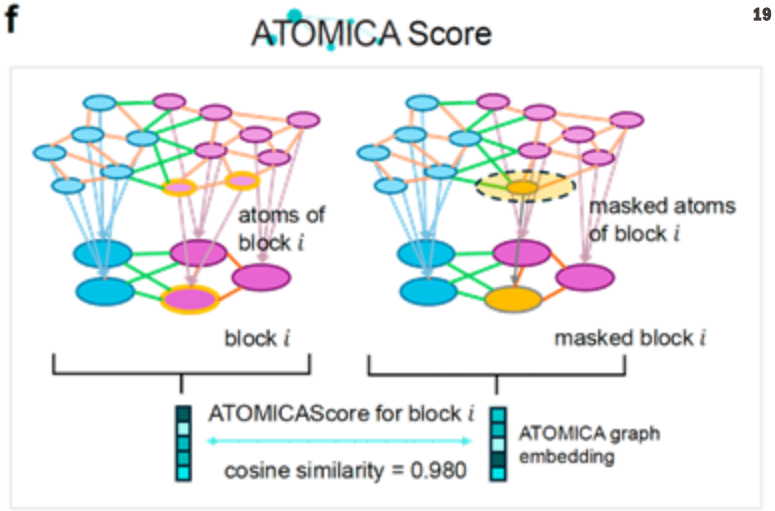

Annotating ligand binding sites

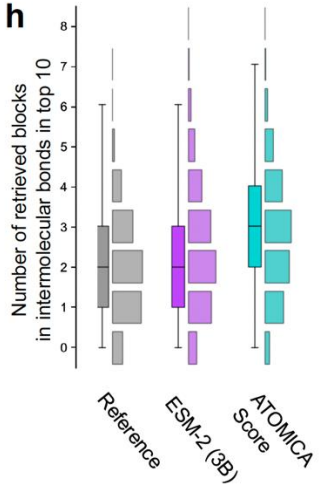



Detection of desease relevant pathways.

# Scoring effects of mutations in binding sites

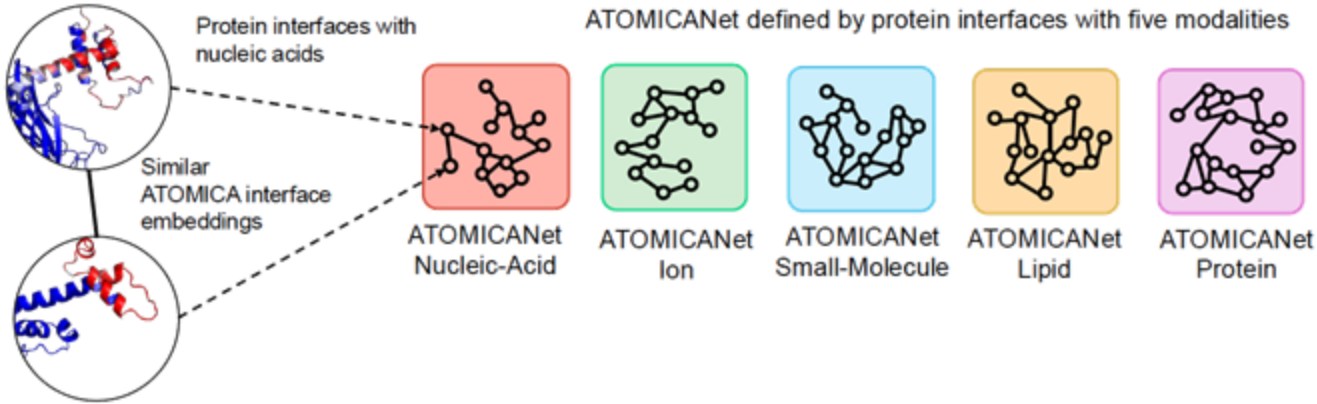
- ATOMICASCORE:
  - $h^{graph}$ embedding of unperturbed graph.
  - $h^{graph'}$ embedding of graph with masked block i.
  - Score = Cos Sim($h^{graph}$, $h^{graph'}$)
- Lowest similarity -> highest importance for interaction.
- Benchmarking recovery of residues essential for binding:
  - Atomica outperformes ESM-2 protein language model.
  - (Modest?) Increase over random selection of key residues from 2 → 2.8/10 recovered residues.



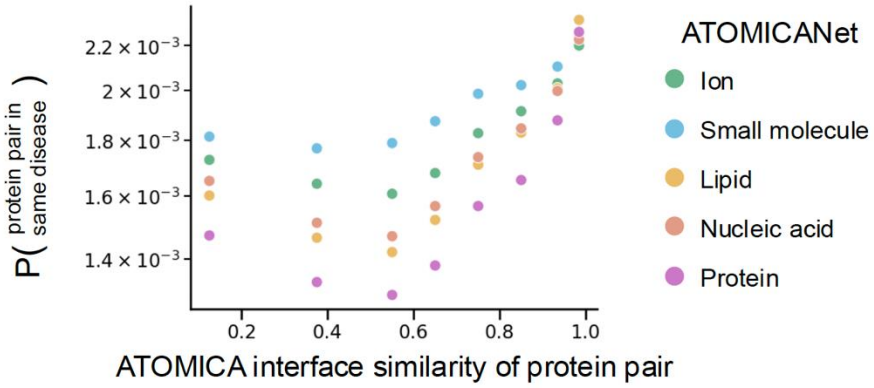
# Constructing interfaceome networks

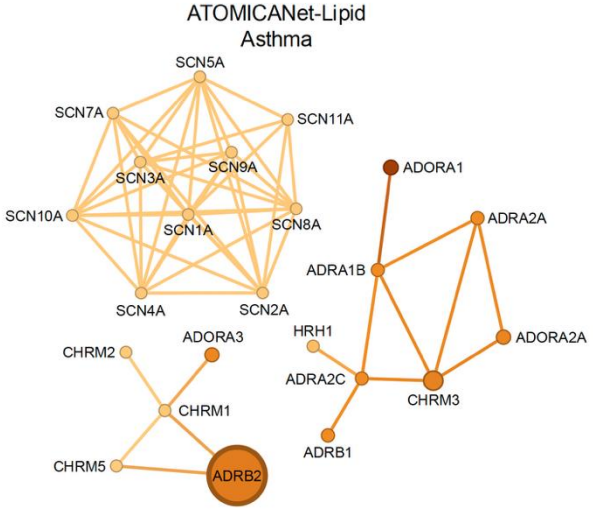
- Atomica-Interface: "Fingerprinting" empty pockets G' to their corresponding complexes G by finetuning ATOMICA-interface $\mathcal{G}$ from a frozen net $\mathcal{H}$ s.t. $\mathcal{H}(\mathrm{G}') \approx \mathcal{G}(\mathrm{G})$.
- Atomica-Net: 23K human proteins in AFDB -> PeSTo ligand binding sites -> Atomica embeddings.
- Separate Atomica-Net variant for all Protein-* networks.
- Network constructed using graph embedding similarities.



# Atomica-Net reconstructs known desease pathways

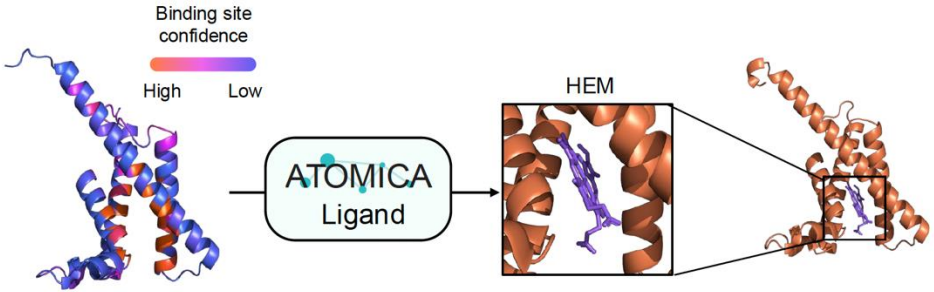
- Co-occurrence of proteins in a desease correlates with ATOMICA interface similarity.
- Proteins in one desease should be found within connected components.
- Case-Study: Atomica-Net Lipid for Asthma-related proteins.
  - Comparing expected largest connected component size vs. actual size in a given desease pathway.
  - For Asthma: Retreived 42 desease-associated proteins.

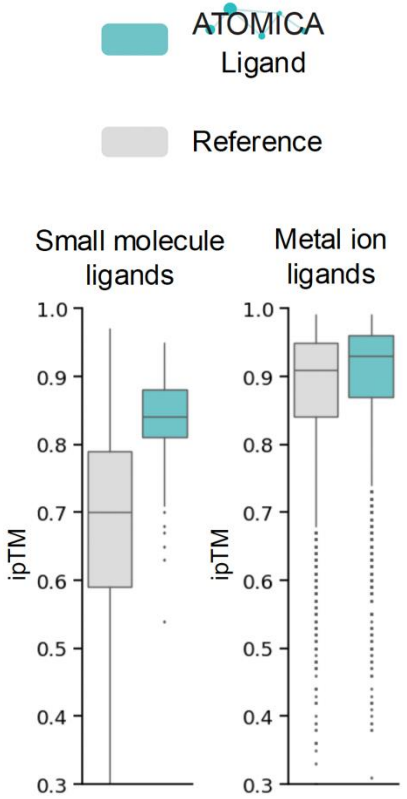


# Ligand binding site annotation with ATOMICA-Ligand

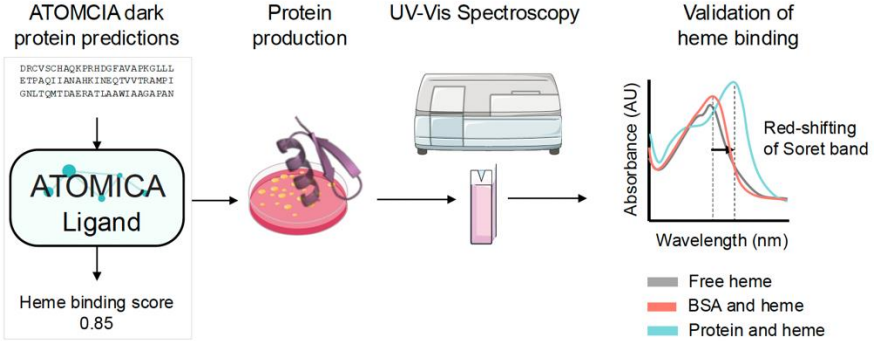
- For each ion and selected small molecules: ATOMICA finetuning to prodict ligand identity for a given pocket.
- Network: Atomica+MPL predicting binary label.
- Application:
  - PeSTo annotation of binding sites in the AFDB, Atomic-Ligand prediction with each finetuned model.
  - Evaluation with AF3 ipTM vs. a reference of randomly selected ligand identity.

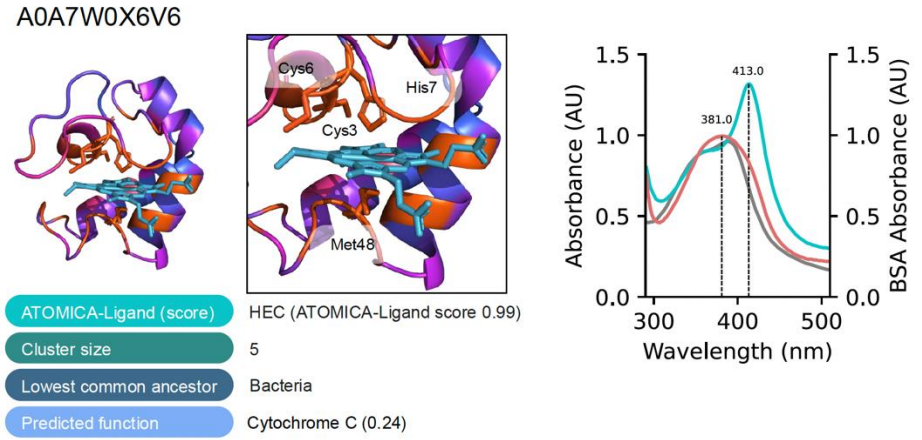


# Identifying functional proteins in the "dark proteome"

- "Dark" clusters in the AFDB: Proteins with low fold-homology to characterized proteins.
- Application of ATOMICA:
  - Identifying Haemoglobin binding pockets.
  - Experimental verification of promising candidates.




Barrio-Hernandez *et al. Nature* 2023

# Discussion

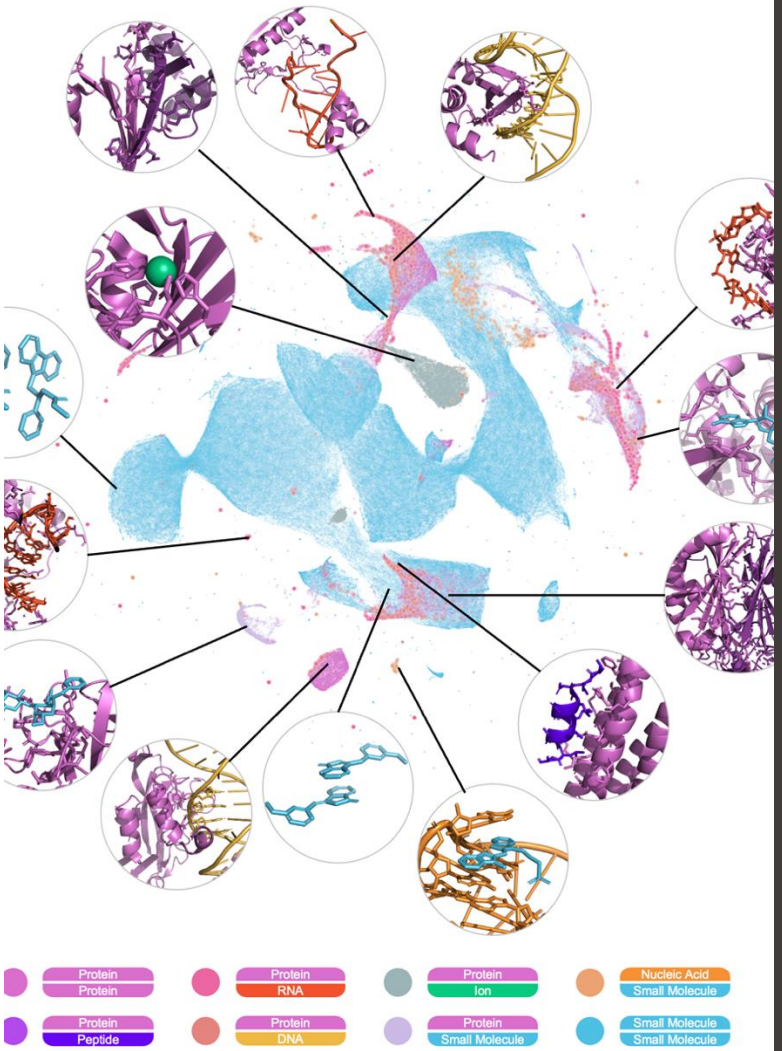
Pros:

- Extensive study with many different promising applications.
- Semantically meaningful representations, as shown by “Embedding Algebra”.
- Generalizability by multimodal pretraining: Accuracy on learning rare interfaces (e.g. Protein-Nucleotide) improves.
- Latent representations improve robustness to apo vs. holo states: Ligand bidning annotation was also possible for apo-like structure predictions.
  - However: Is AF implicitly predicting a holo-state-like structure with the corect ligand?

# Discussion

Cons:

- Reliant on structural data or high-confidence predictions. → Many deseases have at least one intrinsically disordered protein.
- Fine-tuning was be performed for every different application. → Unifying Atomica-Ligand, Atomica-Score, and Atomica-Interface by using different prediction heads possible.
- For desease pathways: The association of Proteins->Pathways alone could also be created with non-structural GWAS, more interesting is the explanation of a potential desease mechanism (not explored in the study).




# Thanks for listening!

Tian will answer questions in person.

For more questions, please drop me an email: adrian.dobbelstein@epfl.ch

# Possible applications in molecular design

Designing small molecules resembling corresponding Protein-Protein-Interactions?

1. Transform a PPI embedding into a corresponding Protein/SM embedding.

2. Steer a SM generative model to resemble an existing PPI (e.g., Feynman-Kac steering)

Alternative: Retraining of the model to be conditioned on ATOMICA embeddings + Embedding similarity loss.