ATOMICA: Learning Universal Representations of Intermolecular Interactions

Anonymous Authors[1]

Abstract

Molecular interactions underlie nearly all biological processes. However, most machine learning models treat molecules in isolation or specialize in a single type of interaction, which prevents generalization across biomolecular classes and limits the ability to systematically model interaction interfaces. We introduce ATOMICA, a geometric deep learning model that learns atomic-scale representations of intermolecular interfaces across diverse biomolecular modalities, including small molecules, metal ions, amino acids, and nucleic acids. ATOMICA uses a self-supervised denoising and masking objective to train on 2,037,972 interaction complexes and generate hierarchical embeddings at the levels of atoms, chemical blocks, and molecular interfaces. The model learns generalizable representations across molecular classes. We apply ATOMICA to the interfaceome and show that proteins that interact similarly with ions, small molecules, nucleic acids, lipids, and proteins tend to be involved in the same disease. We then construct five modality-specific interfaceome networks termed ATOMICANETs, which connect proteins based on interaction interface similarity. These networks identify disease pathways across 27 conditions. Finally, we use ATOMICA to annotate the dark proteome—proteins lacking known structure or function—by predicting 2,646 previously uncharacterized ligand-binding sites for metal ions and cofactors.

1. Introduction

Molecular interactions influence all aspects of chemistry and biology. Despite advances in structure prediction and molecular modeling, prevailing machine learning approaches emphasize modeling molecules in isolation (Rives et al., 2021; Luo et al., 2022) or provide limited modeling of molecular interactions, typically restricted to a specific type of interaction, such as protein-ligand and protein-protein interactions (Gainza et al., 2020). These methods rely on separate architectures for different molecular classes, preventing cross-modality knowledge transfer and limiting the generalizability of learned representations.

Current generative models, including AlphaFold (Google DeepMind AlphaFold Team & Isomorphic Labs Team, 2023) and RosettaFold (Krishna et al., 2024), generate molecular structures but do not explicitly learn representations of intermolecular interactions. We lack a generalizable approach to represent and fingerprint interaction complexes of biomolecules. A universal representation learning model that operates at the atom scale, captures multi-modal molecular interactions, and learns generalizable representations across biomolecular modalities could address this limitation. Existing models primarily learn molecular representations, whereas a model that explicitly represents molecular interactions could unify predictive modeling across different types of biomolecular complexes.

The geometry of intermolecular interactions follows fundamental physical and chemical principles, yet the diversity of interacting molecular types introduces fundamental modeling challenges. Key biomolecular interactions, including protein-RNA, protein-DNA, protein-metal ion, protein-small molecule, protein-peptide, nucleic acid-small molecule, and protein-protein interactions, share universal spatial constraints, such as hydrogen bonding, van der Waals forces, π-stacking, and electrostatic interactions. However, each type of interaction varies dramatically in binding affinity, interface size, chemical composition, and functional roles. For example, protein-protein interfaces are often large (median of 2,000-4,000Å^2), stabilized by multiple hydrophobic and electrostatic contacts, whereas protein-ligand binding sites are smaller (median of 300-1,500Å^2) and shaped by high-affinity, highly specific binding pockets (Evans et al., 2021; Krishna et al., 2024). RNA and DNA interfaces introduce additional sequence-dependent constraints, while metal ions exhibit coordination-specific geometry. Existing deep learning models largely treat these interactions independently, limiting cross-modality knowledge transfer. Jointly modeling atomic interactions across diverse biomolecular interface geometries is essential to developing generalizable, transferable molecular interaction representations, yet remains highly challenging due to the heterogeneity in molecular residues, steric constraints, and biochemical environments.

Present Work. We introduce ATOMICA, an all-atom geo-

metric deep learning model that learns representations of intermolecular complexes across diverse biomolecular modalities, including small molecules, metals, amino acids, and nucleic acids. Unlike existing models that focus on single molecular types, ATOMICA generalizes across modalities by leveraging a pretraining dataset of 2,037,972 interaction complexes. These include 1,747,710 small-molecule interaction complexes from the Cambridge Structural Database (CSD) (Groom et al., 2016) and 290,262 biomolecular complexes from Q-BioLiP and the Protein Data Bank (PDB) (Wei et al., 2023; Yang et al., 2012; Berman et al., 2000). Learning from interactions spanning proteins, nucleic acids, small molecules, and ions enables ATOMICA to generalize across molecular modalities. This cross-domain generalizability improves representation quality in low-data modalities, such as for protein-nucleic acid interactions that are less common in the PDB.

Analysis of the interfaceome reveals that proteins with similar ATOMICA-derived interaction profiles often participate in shared disease pathways across protein interactions with small molecules, ions, lipids, nucleic acids, and proteins. Moving beyond annotated proteins, we apply ATOMICA to the dark proteome—regions of the proteome lacking functional labels (Perdigão et al., 2015; Barrio-Hernandez et al., 2023; Kulkarni & Uversky, 2018). Finetuning ATOMICA enables the annotation of 2,646 binding sites with putative ions and cofactors, revealing functions in ancient and uncharacterized protein families.

2. Related Work

Representation learning for biomolecules. Despite advances in representation learning, existing models remain constrained to specific molecular modalities, limiting their applicability across the biochemical landscape. Protein and nucleic acid models leverage sequence-based tokenization (Lin et al., 2023; Rives et al., 2021; Chen et al., 2022; Boyd et al., 2023; Celaj et al., 2023), whereas small molecules require atomic-scale modeling due to their lack of inherent sequential structure (Chithrananda et al., 2020; Liu et al., 2021a; Zaidi et al., 2022; Atz et al., 2021; Wang et al., 2022b; Fang et al., 2022).

Predictive models for molecular interactions. Current molecular interaction models are specialized, with distinct architectures designed for protein-ligand binding affinity (Moesser et al., 2022; Yan et al., 2023; Moon et al., 2022; Li et al., 2021; Meng & Xia, 2021), binding site prediction (Meller et al., 2023; Krapp et al., 2023; Jiménez et al., 2017; Kandel et al., 2021), protein-peptide interactions (Tsaban et al., 2022; Cunningham et al., 2020; Lei et al., 2021), protein-protein interactions (Gainza et al., 2020; Sverrisson et al., 2021; Gainza et al., 2023; Bryant et al., 2022; Das & Chakrabarti, 2021; Renaud et al., 2021), and protein-RNA recognition (Lam et al., 2019; Xia et al., 2021; Alipanahi et al., 2015; Wei et al., 2022; Sun et al., 2021; Rube et al., 2022). This siloed approach prevents knowledge transfer across molecular classes even though interactions between proteins, nucleic acids, small molecules, and ions obey shared physicochemical principles.

Universal generative models for biomolecular structure prediction. Structure-based generative models have demonstrated the feasibility of learning across all biomolecular modalities present in the Protein Data Bank (Krishna et al., 2024; Google DeepMind AlphaFold Team & Isomorphic Labs Team, 2023). However, existing approaches do not yet unify molecular representations across interaction types, leaving open the question of whether a single model can capture the full spectrum of biomolecular interactions.

3. ATOMICA Approach

We model the interactions between molecules, which is contrary to prior work focused on modeling individual molecules or protein surfaces. By modeling intermolecular interactions universally across all modalities, we instill the inductive prior that they are all fundamentally governed by the same chemistry principles of intermolecular bonding, such as hydrogen bonding, hydrophobic interactions, and Van der Waals forces.

3.1. Problem Setup: Self-Supervised Learning on Interaction Complexes

Given is an unlabeled pretraining dataset of graphs of molecular complexes, $\mathcal{D} = \{G^i \mid i = 1, \ldots, N\}$, and a target dataset of labeled graphs of molecular complexes $\mathcal{S} = \{(G^i_{\text{target}}, y_i) \mid i = 1, \ldots, M\}$, where $M << N$. Our goal is to pretrain a model $\mathcal{F}$ on $\mathcal{D}$ such that it generates representations $\mathbf{h}_i = \mathcal{F}(G^i)$ for every intermolecular patch G^i that are chemically informative, and $\mathcal{F}$ can also be finetuned on $\mathcal{S}$ to predict y_i for every G^i_{target}.

3.2. Overview of ATOMICA Model Architecture

Hierarchical graph input. We represent each interaction complex using a hierarchical graph that models both atomic-level details and higher-order chemical structure (Fig. 1). At the first level, nodes represent individual atoms, each defined by its element type and 3D spatial coordinates. At the second level, we group atoms into chemically meaningful blocks—such as amino acids in proteins, nucleotides in nucleic acids, or functional moieties in small molecules—and construct a block-level graph (Hermosilla et al., 2021; Wang et al., 2022a; Kong et al., 2023). This hierarchical design captures both local atomic interactions and broader structural organization and has theoretically higher expressive power than purely atom-level graphs (Wollschläger et al.,

2024). Within each level, we define two types of edges: intramolecular edges connect nearby nodes within the same molecule, and intermolecular edges connect nearby nodes across the interface between two interacting molecules (details are available in Appendix A).

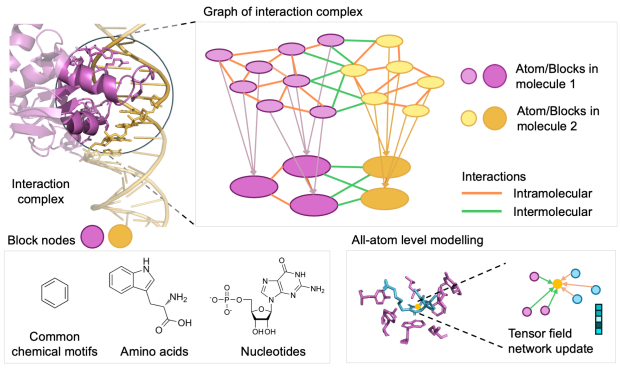

Figure 1. Overview of ATOMICA architecture, interaction complexes are modeled at the atom and block level. Message passing between nodes at each level is done via intermolecular and intramolecular edges.

ATOMICA equivariant all-atom graph neural network. ATOMICA is a self-supervised geometric graph neural network that learns multi-scale embeddings at the atom, block, and graph level from the structure of interacting two molecules (Fig. 1). Unlike modality-specific models, ATOMICA is capable of generating embeddings at the interface for any complex of interacting molecular modalities (small molecules, metals, amino acids, and nucleic acids). We use SE(3)-equivariant tensor field networks for message passing (Appendix B), which have been used to predict interatomic potentials (Batzner et al., 2022; Musaelian et al., 2023), molecular coupling (Corso et al., 2023), and scoring RNA structure (Townshend et al., 2021). Message passing is first done at the atom-level across intermolecular and intramolecular edges and it is then pooled to the blocks the nodes belong to. Message passing is completed again at the block-level and graph-level embeddings are then produced by pooling the block-level embeddings.

Self-supervised learning with ATOMICA. To learn high-quality representations, we employ a denoising and masked block strategy (Appendix C). Denoising is effective as a pretraining objective to learn representations of 3D conformations of single molecules for property prediction (Luo et al., 2022; Zaidi et al., 2023; Zhou et al., 2023; Godwin et al., 2022), unsupervised binding affinity prediction (Jin et al., 2023). Masking is a powerful self-supervised objective for learning representations of protein sequences (Rives et al., 2021) and nucleic acid sequences (Dalla-Torre et al., 2025). The ATOMICA pretraining strategy applies a rigid SE(3) transformation as well as random rotation and torsion angles of one of the molecular entities at the interface. The model denoising output is optimized to minimize the distance to the score function of the global translation, global rotation, and torsion noise distributions (Corso et al., 2023; Jin et al., 2023). By denoising and masking one molecular interface with respect to the other, this approach aims to capture the chemical, structural, and geometric patterns of intermolecular interaction.

Formally, given $G^i \in \mathcal{D}$, which is comprised of atom and block nodes from two interacting molecules. $\widetilde{G}^i$ is a perturbed graph created by applying two transformations to a molecule in G^i, which is selected uniform randomly:

- Rigid rotation and translation: A rotation vector is sampled $\boldsymbol{\omega} \sim p(\boldsymbol{\omega}) = \mathcal{N}_{SO(3)}$ and we apply the rotation of all atom and block coordinates about the center of the selected molecule. A translation vector is sampled $\mathbf{t} \sim p(\mathbf{t}) = \mathcal{N}(0, \sigma_t^2 \mathbf{I})$ and we apply this translation to all atom and block coordinates of the selected molecule.
- Torsion angle noising: Torsion angles are sampled $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}) = \mathcal{N}_{SO(2)^m}$ where m is the number of rotatable bonds in the molecule. For peptides, proteins, RNA and DNA we only perturb rotatable bonds in the side chain.

ATOMICA is pre-trained to predict the rotation score $\mathbf{s}_{\boldsymbol{\omega}} \in \mathbb{R}^3$, translation score $\mathbf{s}_{\mathbf{t}} \in \mathbb{R}^3$, and torsion score $\mathbf{s}_{\boldsymbol{\theta}} \in \mathbb{R}^m$ from $\widetilde{G}^i$ (details are available in Appendix C).

In addition to denoising, ATOMICA is also pretrained with identifying masked out block identities. For each graph G^i, 10% of blocks are randomly sampled and their block identities are replaced with the special 'mask' block and we denote these blocks as $\mathcal{B}$.

3.3. Pretraining Dataset

We assembled a dataset of pairs of interacting molecular entities from the Cambridge Structural Database (CSD) v2022.3.0 (Groom et al., 2016) and Q-BioLiP (Wei et al., 2023; Yang et al., 2012) which includes all biologically relevant intermolecular interactions across all modalities available in the Protein Data Bank. This results in 1,767,710 interacting pairs of small molecules. From Q-BioLiP, which includes structures of protein complexes with proteins, DNA, RNA, peptides, ligands, and ions, as well as nucleic acid ligand structures from the PDB, we obtain 337,993 interaction complexes. The interaction interface between two entities is defined by atoms within an 8 Å distance to the other molecule. For larger molecules (proteins and nucleic acids), we crop the molecules to only keep residues at the interaction interface. Details are available in Appendix D.

4. Experiments

4.1. Latent space of ATOMICA captures physicochemical features

Experimental Setup. ATOMICA is used to generate multi-scale embeddings at the atom, block, and graph level for complexes of molecules between ions, small molecules, peptides, proteins, and nucleic acids. We use uniform manifold approximation and projection (UMAP) to project the latent space of graph embeddings of all interaction complexes in our dataset (Fig. 2). We also explore the mean node embeddings of all nodes of each atomic element and block types in the pretraining dataset by projecting into 2D with principal component analysis (PCA).

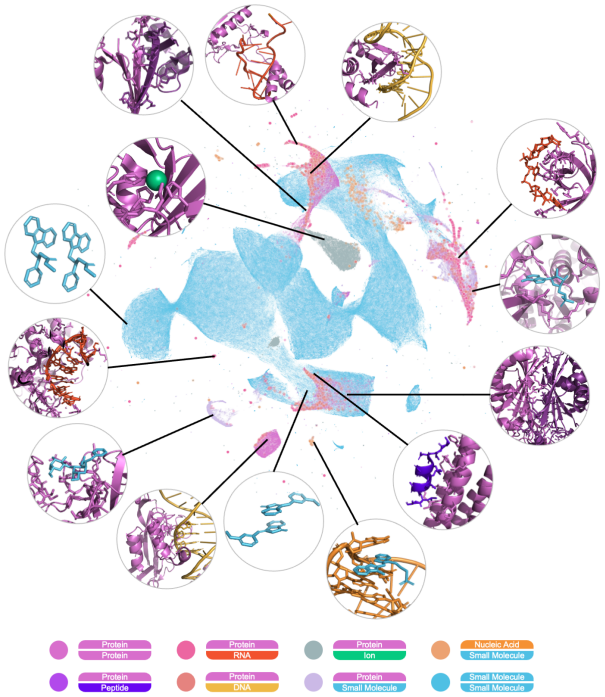

Figure 2. UMAP of latent space of all interaction complexes seen during pretraining of ATOMICA.

Results. While not explicitly provided by the model, interactions are enriched by the modalities of interactions and capture relative chemical similarity of interactions. For example, the embeddings of protein-protein interactions lie closer to the embeddings of protein-peptide interactions. We confirm this by comparing the distribution of cosine similarity between protein-protein and protein-peptide graph embeddings compared to the cosine similarity between any ATOMICA embedding and protein-peptide graph embeddings (KS-statistic = 0.100, p-value < 0.001). We also observe the organization of the latent space of the mean node embeddings according to periodic table location (Fig. 3a). Additionally, we see organization according to the chemical properties of the amino acids (Fig. 3b) and these are distinct from DNA and RNA nucleotides. This suggests ATOMICA has learned chemical similarity from the patterns of occurrence of elements and amino acids observed during pretraining.

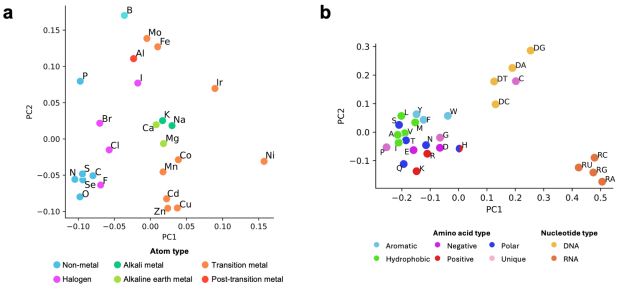

Figure 3. **a** Principle components of mean embedding of elements in the pretraining validation and test set. **b** Principle components of mean embedding of amino acids and nucleotides in the pretraining validation and test set.

4.2. ATOMICA identifies residues on interaction interfaces involved in intermolecular bonds

Experimental Setup. We examine the ability of ATOMICA to differentiate between blocks that are involved in intermolecular bonds (hydrogen bonding, pi-stacking, hydrophobic interactions) and other blocks present at the interface (Fig. 4a). To identify amino acids involved in intermolecular interactions, we analyzed protein-small molecule complexes in the test set using PLIP (Adasme et al., 2021). We then quantified the contribution of each amino acid to the interaction by calculating an importance score, ATOMICASCORE (Fig. 4b). For amino acid i, ATOMICASCORE is defined as $a_i = \text{sim}\left(\mathbf{h}_G, \mathbf{h}_{G_i^{\text{mask}}}\right)$, where $\mathbf{h}_G$ is the embedding of the original complex, and $\mathbf{h}_{G_i^{\text{mask}}}$ represents the embedding of the modified complex in which amino acid i has been replaced with a special mask token and its constituent atoms substituted with a single special mask atom. Here, sim denotes the cosine similarity between the two embedding vectors. In total, we analyze 5,691 protein-small molecule complexes with at least 20 amino acid blocks at the interface. We report precision at rank 10 (Fig. 4c). For reference, we compare this to randomly nominating 10 amino acids at the interface, and to ESM-2 (3 billion parameters) (Rives et al., 2021) and score amino acids by the log likelihood of the masked mutant compared to the original amino acid (Meier et al., 2021) (Fig. 4d). Although ATOMICA is not trained on any labels related to intermolecular bonds, this setup enables a zero-shot probing of its learned representations.

Results. We evaluate precision at rank 10 and find that ATOMICASCORE achieves the highest performance, with an average of 2.7 amino acid blocks in intermolecular bonds

Table 1. AUPRC performance of ATOMICA on masked-block prediction: models pretrained on *all* interacting-modality pairs vs. single-pair baselines.

Modality	All pairs		One pair	
	Mean	Std	Mean	Std
SM-SM	0.958	0.006	0.958	0.003
Protein-protein	0.789	0.002	0.774	0.002
D/RNA-SM	0.758	0.060	0.595	0.034
Protein-DNA	0.707	0.014	0.243	0.007
Protein-peptide	0.666	0.004	0.322	0.006
Protein-ion	0.621	0.020	0.540	0.021
Protein-RNA	0.552	0.008	0.187	0.008
Protein-SM	0.331	0.005	0.307	0.005

SM = small molecule

retrieved, followed by ESM-2 with 2.4, and the random reference with 2.0 (Fig. 4d). ATOMICA's ability to identify intermolecular bonds is also uniformly true across the most common types of intermolecular bonds observed (Fig. S1).

4.3. Pretraining on interaction complexes of multiple modalities leads to better generalizability

Experimental Setup. To evaluate the benefits of incorporating multiple molecular modalities in pretraining ATOMICA, we compare the full ATOMICA model against identical model architectures that were each pretrained exclusively on individual pairs of interacting molecular modalities (Fig. 5a). We reserve a test set of interface complexes with a maximum of 30% sequence similarity and minimal small molecular structure similarity to structures observed in training and validation to evaluate the models on. To evaluate generalizability and model embedding quality, we use the accuracy of the models on masked block identity prediction. This task tests whether the model can recover missing structural components based on their context, which reflects chemically meaningful constraints such as conserved motifs or binding site configurations.

Results. Pretraining across molecular modalities of interacting molecules improves block identity recovery over pretraining on a singular type of interacting molecules on the test set for all pairs of modalities (Table 1). ATOMICA's performance gains are correlated with dataset size (Fig. 5b), reflecting established scaling laws in LLMs where performance improves with dataset size (Kaplan et al., 2020). ATOMICA also demonstrates for the first time an approach to address the limited availability of structural data for interaction complexes with DNA, RNA, and peptide modalities.

4.4. Proteins that share similar ATOMICA protein interfaces tend to be involved in the same disease

Complex diseases are caused by a signaling network's dysregulation rather than a single protein (Menche et al., 2015) and often involve dysfunctional interactions with ions (Leal et al., 2012), small molecules (Sawicki et al., 2015; Shan et al., 2015), lipids (Saliba et al., 2015), or nucleic acids (Tateishi-Karimata & Sugimoto, 2021). Proteins involved in the same disease tend to be clustered in the same network neighborhood where network relations are defined via protein-protein interactions and maps of cellular pathways (Kratz et al., 2023; Zheng et al., 2021). In this section, we test the hypothesis if relations defined by similar interactions with other molecular modalities, given by ATOMICA embeddings, are likely to be involved in the same disease.

Experimental setup. To test this hypothesis, we first embed the human interfaceome, defined as the set of human protein interfaces that mediate interactions with other molecules, including ions, small molecules, nucleic acids, lipids, and proteins. We use PeSTo (Krapp et al., 2023) to predict modality-specific binding sites for 23,391 protein structures predicted by AlphaFold2. We finetune ATOMICA-Interface from ATOMICA to support embedding of protein interfaces instead of complexes (details in Appendix F). To ensure we are working with high-quality protein structures, we remove binding sites of protein interfaces that have low confidence pLDDT scores < 70. For disease proteins, we select a diverse set of 82 diseases and their disease-associated proteins from OpenTargets (Buniello et al., 2025).

Results. We confirm that proteins with similar interaction profiles in the interfacome networks often participate in the same disease pathways, as the probability of pairs of nodes being involved in the same disease is higher if the two nodes have higher ATOMICA similarity (Fig. 6). We next explore how ATOMICA similarity across the interfaceome can be applied to studying disease proteins in Section 4.5.

4.5. Disease pathways in interfaceome-based ATOMICANET networks

Formally, for disease d with associated proteins V_d, the disease pathway $H_d = (V_d, E_d)$ is a subnetwork of the network of proteins. Since interactions of proteins with other molecular modalities are often implicated in diseases, relying on protein-protein interactions and maps of cellular pathways may fail to capture this information. We study disease pathways from a new angle with ATOMICA to with ATOMICANETs where proteins implicated in the same disease are more likely to share similar ion, small molecule, nucleic acid, and lipid interactions.

Construction of ATOMICANET. All binding sites for each modality extracted with PeSTo are embedded with ATOMICA-Interface. We compute pairwise cosine similarity matrices from the embeddings for each of the three ATOMICA-Interface replicates and then average them to produce a single, consolidated cosine similarity matrix. Using the resultant cosine similarity matrix, we then construct

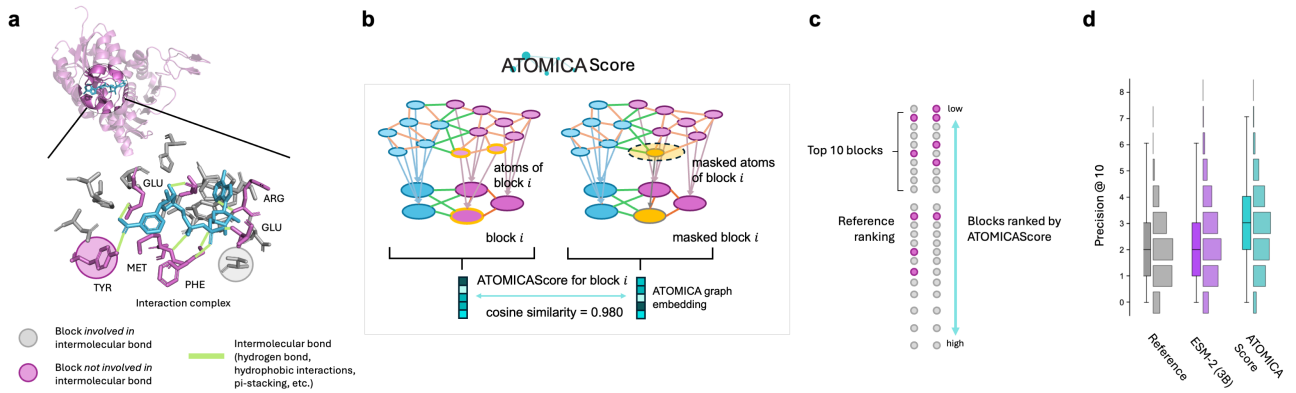

Figure 4. **a** Some blocks in the interaction complex are involved in intermolecular bonds (hydrogen bonds, hydrophobic interactions, pi-stacking, etc.). **b** Definition of ATOMICASCORE for block i, cosine similarity between interaction graph with block i masked and unmasked. **c** Nomination of blocks involved in intermolecular bonds at interfaces based on ATOMICASCORE. The reference ranking is determined by random ordering of the blocks at the interaction interface. **d** Number of blocks involved in intermolecular bonds in the top 10 nominated blocks of ATOMICASCORE, ESM-2 (3B parameters), and reference for protein-small molecule complexes in the pretraining test set.

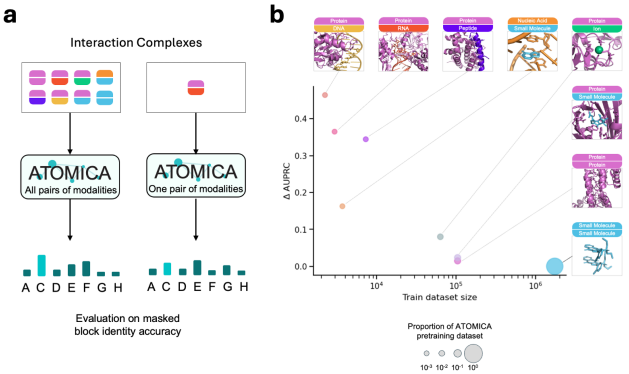

Figure 5. **a** Schema to test generalizability of representations learned by ATOMICA trained on all pairs of modalities compared to trained on one pair of modalities. We evaluate quality of representations based on masked block identity accuracy. **b** Increase in AUPRC between models trained pretrained on all pairs v.s. one pair.

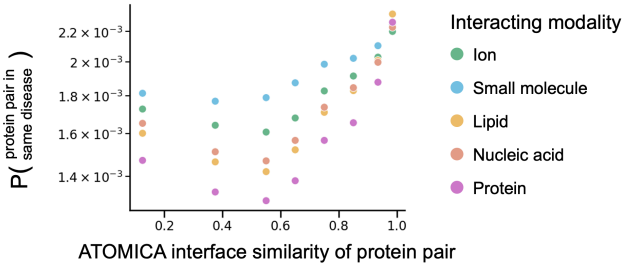

Figure 6. Cosine similarity of ATOMICA embeddings of protein interface pairs across five interacting modalities compared to the probability of the protein pair being involved in the same disease.

a network for each modality based on a cosine similarity threshold and enforce that each node in the network has a maximum degree of 50. Cutoffs are defined such that 90% of the proteins in each modality are in the largest connected component. We construct these networks using NetworkX (Hagberg et al., 2008). In total, for the largest connected component in each network, we have 5,831 nodes in ATOMICANET-Ion, 5,246 nodes in ATOMICANET-Small-Molecule, 5,974 nodes in ATOMICANET-Nucleic-Acid, 6,055 nodes in ATOMICANET-Lipid, and 15,450 nodes in ATOMICANET-Protein (Fig. 7a). Visualisations of the networks are constructed with Gephi (Bastian et al., 2009).

Observation of disease pathways in ATOMICANETs. For the target-disease associations, we study their disease pathways across the five ATOMICANETs. A disease pathway is one or more connected subgraphs comprised of disease proteins (Menche et al., 2015), with a minimum requirement of 25 associated genes for a disease for there to be an observable disease pathway. We refer to a disease d with associated proteins in a modality network V_d^{modality} and the disease pathway is the undirected subgraph $H_d^{\text{modality}} = \left(V_d^{\text{modality}}, E_d^{\text{modality}}\right)$. Following (Agrawal et al., 2018), we use their definition of the size of the largest pathway component as the fraction of disease proteins that lie in H_d^{modality}'s largest connected component. For all modalities with $|V_d^{\text{modality}}| > 25$, we analyze the size of the largest pathway component. To assess the statistical significance of the observed pathway size, we compared it against a distribution derived from 1,000 randomized sets of disease proteins. These randomized sets were constructed to match the degree distribution of the

original disease proteins, thereby accounting for the heterogeneous connectivity patterns in ATOMICANETs. For each network, we applied the Benjamini-Hochberg procedure to correct for multiple hypothesis testing, considering results with adjusted p-values < 0.05 as statistically significant. Across the five networks, the average size of the largest pathway component for ATOMICANET-Ion is 11%, for ATOMICANET-Small-Molecule is 11%, for ATOMICANET-Lipid is 16%, for ATOMICANET-Nucleic-Acid is 10%, and for ATOMICANET-Protein is 6% Fig. 7c). In the following section, we highlight some of the largest pathway components observed for diseases across the ATOMICANET-Ion, Small-Molecule, and Lipid.

Examining disease pathways in ATOMICANETs. First we look at disease pathways analyzed in ATOMICANET-Lipid, of the 40 diseases with sufficient disease proteins we found that 22 diseases exhibited significantly larger largest pathway components than expected, and 11 diseases had significantly fewer disconnected pathway components than expected. Asthma has 43 disease proteins in ATOMICANET-Lipid (Fig 7b), and has a well-observed disease pathway (p-value < 0.001 for size of largest pathway component and p-value < 0.001 for number of pathway components) with 10 proteins in the largest pathway component, which is comprised of sodium channel family proteins (OpenTargets mean strength of evidence = 0.54, mean evidence sources = 5.2). In the second and third largest pathway component, we observe 8 and 5 proteins, respectively, both involving G protein-coupled receptors (adenosine, α/β-adrenergic, muscarinic, and histamine receptors). These clusters have a mean strength of evidence of 0.61 and 0.56 with on average 66 and 245 sources of evidence (Ochoa et al., 2021). Proteins in these two components form key interactions with PIP2, a minority lipid component of the cell membrane (Yen et al., 2018a).

Next in ATOMICANET-Ion, 10 diseases had significantly larger pathway components than expected and 11 diseases had fewer disconnected pathway components than expected of the 35 diseases with sufficient disease proteins. Myeloid leukemia has 53 disease proteins in ATOMICANET-Ion (Fig 7b), with 12 proteins in the largest pathway component (p-value < 0.001) with a mean strength of evidence of 0.60 and on average 401 sources of evidence per disease protein (Ochoa et al., 2021). This component includes TET2, a Fe^{2+} binder, which plays a key role in active DNA demethylation and is frequently mutated in acute myeloid leukemia (Yen et al., 2018b). Four DNA binding proteins with zinc finger domains are also observed (DNMT1, POLE, WT1, PHF6) in the largest pathway component on ATOMICANET-Ion, showing the ability of ATOMICANET to capture disease-relevant similar interaction patterns. Other proteins in this cluster include isoforms of protein kinase C and serine/threonine-protein kinase D proteins.

In ATOMICANET-Small-Molecule, one disease had a significantly larger pathway component than expected of the 37 diseases with sufficient disease proteins. Hypertrophic cardiomyopathy has 45 associated proteins in ATOMICANET-Small-Molecule (Fig 7b), and the largest pathway component is of size 7 (p-value $= 0.037$) with a mean strength of evidence of 0.70 and on average 630 sources of evidence per disease protein (Ochoa et al., 2021). Since ATOMICANET-Small-Molecule connects proteins that share similar binding sites, we find that the proteins in this component share nucleotide (GTP/GDP, ATP/ADP) binding sites. These proteins include: myosin heavy chain proteins (MYH6, MYH7B) – which are responsible for force generation in cardiac muscle and are frequently mutated in patients with hypertrophic cardiomyopathy (Jiang et al., 2013; McNally, 2002), cardiac actin (ACTC1) – a crucial sarcomeric protein which is also strongly associated with the disease (Despond & Dawson, 2018), and HRAS – a GTPase regulating a host of signaling pathways and cellular responses (Matsuda et al., 2017).

For ATOMICANET-Nucleic-Acid and ATOMICANET-Protein we do not observe any statistically significant pathway components. These networks also have relatively smaller largest pathway components with a mean size of 3.4 members for ATOMICANET-Nucleic-Acid and 6.0 for ATOMICANET-Protein, compared to 7.3 for ATOMICANET-Small-Molecule, 8.3 for ATOMICANET-Ion, and 8.9 for ATOMICANET-Lipid. Thus, disease pathways are likely currently unobservable in ATOMICANET-Nucleic-Acid and ATOMICANET-Protein (Menche et al., 2015).

4.6. Ligand annotation for binding sites in the dark proteome

In this section we demonstrate the application of ATOMICA outside of the human proteome, and to poorly annotated proteins. Ion and cofactor binding sites are conserved functional features widely distributed throughout the proteome (Cammisa et al., 2013; Harel et al., 2014). Even among proteins that differ significantly in their overall sequence or structure, ligand-binding regions often align with functional domains preserved across evolution (Harel et al., 2014). Structural databases such as PDB document compelling examples where proteins with minimal sequence similarity (as low as 22%) still maintain strikingly similar ligand-binding sites (Corso et al., 2024). Suggesting the potential of structure-based approaches, such as ATOMICA, to annotate functional binding sites in regions of the proteome that currently lack any functional description, collectively known as the *dark proteome* (Perdigão et al., 2015; Barrio-Hernandez et al., 2023; Kulkarni & Uversky, 2018). Currently, 711,705 protein clusters (30.9% of all clusters cataloged in the AlphaFold (AFDB) FoldSeek) fall into this dark category. Investigating these dark clusters pro-

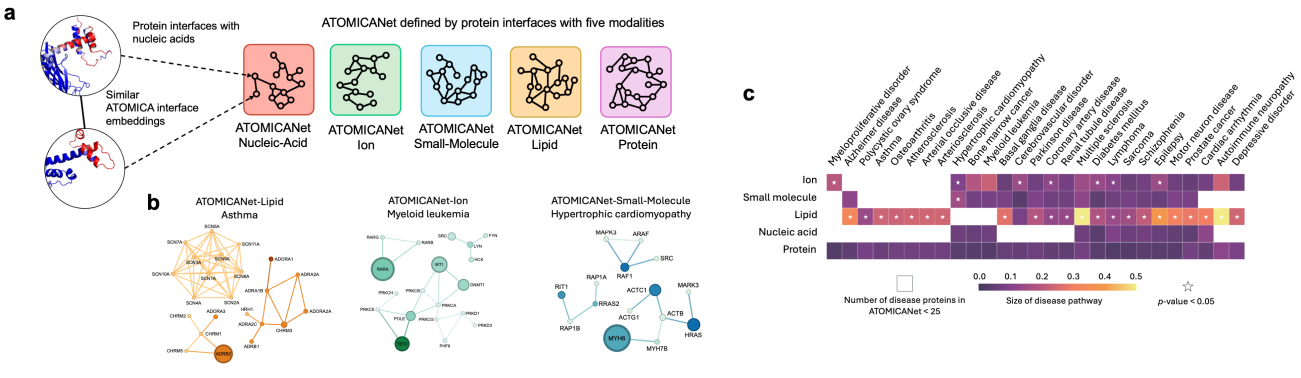

Figure 7. **Interfaceome disease pathways on ATOMICANETs.** **a** Set up of the modality specific networks based on ATOMICA embedding similarity of protein interfaces with ions, small molecules, lipids, nucleic acids, and proteins. **b** The three largest pathway components for: asthma in ATOMICANET-Lipid, myeloid leukemia in ATOMICANET-Ion, hypertrophic cardiomyopathy in ATOMICANET-Small-Molecule. **c** Relative size of largest pathway component across diseases for each modality network. We display only the diseases which have statistically larger pathway components than expected in at least one ATOMICANET modality.

vides opportunities to discover previously unknown protein functions, elucidate novel molecular mechanisms, and reconstruct evolutionary trajectories leading to present-day protein diversity (Perdigão et al., 2015; Bitard-Feildel & Callebaut, 2017; Levitt, 2009).

Experimental Setup. Dark proteins often differ significantly from characterized proteins in both sequence and structure, sequence-based methods have difficulty assigning functional annotations such as ligand or cofactor binding (Hekkelman et al., 2023). ATOMICA addresses this challenge by structurally annotating ion and cofactor binding sites in the dark proteome. We restrict our analysis to dark clusters with high confidence AlphaFold2 structures and identify ligand binding sites are on the surface of proteins with PeSTo. In total, 2,851 proteins are identified with ion binding sites and 969 proteins are identified with small molecule binding sites. We finetune ATOMICA to predict ion and cofactor identities given respective protein pockets from structures in the PDB for 9 metal ions and 12 commonly found cofactors (Appendix G).

Results. ATOMICA annotates metal ion binding sites for 2,565 out of 2,851 proteins and ligand binding sites for 81 out of 969 proteins. Using AlphaFold3 we confirm the quality of ATOMICA predictions with ipTM scores of the complexes, which serve as a quantitative metric for the generation quality of complexes (Bhat et al., 2023; Abramson et al., 2024). The results from ATOMICA are statistically significantly higher than reference complexes for ions (KS Statistic: 0.11, p-value < 0.001) and ligands (KS Statistic: 0.54, p-value < 0.001) (Fig. 8). Reference complexes are determined by randomly assigning ions and ligands to the predicted binding sites in the dark proteome. These annotations span proteins across cellular organisms with a total of 1,265 unique species, of which 1,051 are Bacteria, 99 are Eukaryota, and 115 are Archaea.

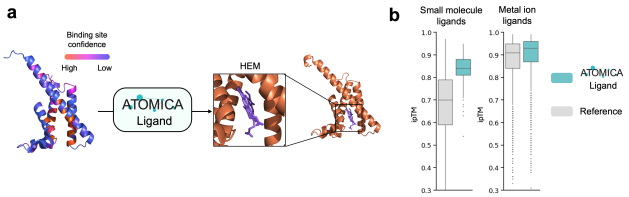

Figure 8. **a** Prediction of ligands for metal ion and small molecule binding sites of proteins in the dark proteome. **b** AlphaFold3 ipTM scores of complexes from ATOMICA-Ligand annotated small molecule and metal ion compared to reference.

5. Conclusion

ATOMICA is a model for representing intermolecular interactions across molecular modalities. By pretraining on over two million molecular complexes involving small molecules, metal ions, amino acids, and nucleic acids, ATOMICA learns hierarchical, chemically grounded embeddings that generalize across interaction types. Exploring the human interfacome with ATOMICA embeddings also shows proteins sharing similar interaction interfaces are likely to be involved in the same disease. ATOMICA generalizes to previously uncharacterized proteins in the dark proteome, allowing the annotation of ion and cofactor binding sites in structurally and functionally novel protein families. As efforts toward a virtual cell intensify (Bunne et al., 2024), approaches such as ATOMICA will be necessary to model the full spectrum of molecular interactions, including protein-ion, protein-small molecule, protein-protein, and protein-nucleic acid contacts, at atomic resolution. Extending ATOMICA to integrate sequence-derived and experimental interaction evidence will improve its applicability in capturing molecular interactions in biological systems.

Impact Statement

This work advances the field of machine learning by introducing a universal representation learning framework for modeling intermolecular interactions across diverse biomolecular modalities. By generalizing across molecular modalities, it may accelerate biomedical research and therapeutic development. While the model could be applied in settings with dual-use potential, such as compound design, we rely on public datasets and focus on medically relevant applications.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pp. 1–3, 2024.

Adasme, M. F., Linnemann, K. L., Bolz, S. N., Kaiser, F., Salentin, S., Haupt, V. J., and Schroeder, M. Plip 2021: Expanding the scope of the protein–ligand interaction profiler to dna and rna. *Nucleic acids research*, 49(W1): W530–W534, 2021.

Agrawal, M., Zitnik, M., and Leskovec, J. Large-scale analysis of disease pathways in the human interactome. In *PACIFIC SYMPOSIUM on BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*, pp. 111–122. World Scientific, 2018.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019.

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

Atz, K., Grisoni, F., and Schneider, G. Geometric deep learning on molecular representations. *Nature Machine Intelligence*, pp. 1–10, 2021.

Bajusz, D., Rácz, A., and Héberger, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):1–13, 2015.

Barrio-Hernandez, I., Yeo, J., Jänes, J., Mirdita, M., Gilchrist, C. L., Wein, T., Varadi, M., Velankar, S., Beltrao, P., and Steinegger, M. Clustering predicted structures at the scale of the known protein universe. *Nature*, 622(7983):637–645, 2023.

Bastian, M., Heymann, S., and Jacomy, M. Gephi: An open source software for exploring and manipulating networks, 2009. URL http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., and Kozinsky, B. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

Bhat, S., Palepu, K., Hong, L., Mao, J., Ye, T., Iyer, R., Zhao, L., Chen, T., Vincoff, S., Watson, R., et al. De novo design of peptide binders to conformationally diverse targets with contrastive language modeling. *bioRxiv*, pp. 2023–06, 2023.

Biewald, L. Experiment tracking with weights and biases, 2020. URL https://www.wandb.com/. Software available from wandb.com.

Bitard-Feildel, T. and Callebaut, I. Exploring the dark foldable proteome by considering hydrophobic amino acids topology. *Scientific reports*, 7(1):41425, 2017.

Boyd, N., Anderson, B. M., Townshend, B., Chow, R., Stephens, C. J., Rangan, R., Kaplan, M., Corley, M., Tambe, A., Ido, Y., et al. Atom-1: A foundation model for rna structure and function built on chemical mapping data. *bioRxiv*, pp. 2023–12, 2023.

Bryant, P., Pozzati, G., and Elofsson, A. Improved prediction of protein-protein interactions using alphafold2. *Nature communications*, 13(1):1265, 2022.

Buniello, A., Suveges, D., Cruz-Castillo, C., Llinares, M. B., Cornu, H., Lopez, I., Tsukanov, K., Roldán-Romero, J. M., Mehta, C., Fumis, L., et al. Open targets platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Research*, 53(D1):D1467–D1475, 2025.

Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., Alexandrov, T., AlQuraishi, M., Brennan, P., Burkhardt, D. B., et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024.

Cammisa, M., Correra, A., Andreotti, G., and Cubellis, M. V. Identification and analysis of conserved pockets on protein surfaces. *BMC bioinformatics*, 14:1–9, 2013.

Celaj, A., Gao, A. J., Lau, T. T., Holgersen, E. M., Lo, A., Lodaya, V., Cole, C. B., Denroche, R. E., Spickett, C., Wagih, O., et al. An rna foundation model enables discovery of disease mechanisms and candidate therapeutics. *bioRxiv*, pp. 2023–09, 2023.

Chen, J., Hu, Z., Sun, S., Tan, Q., Wang, Y., Yu, Q., Zong, L., Hong, L., Xiao, J., Shen, T., et al. Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *arXiv preprint arXiv:2204.00300*, 2022.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607. PMLR, 2020.

Chithrananda, S., Grand, G., and Ramsundar, B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking. 2023.

Corso, G., Deng, A., Fry, B., Polizzi, N., Barzilay, R., and Jaakkola, T. Deep confident steps to new pockets: Strategies for docking generalization. *ArXiv*, pp. arXiv–2402, 2024.

Cunningham, J. M., Koytiger, G., Sorger, P. K., and AlQuraishi, M. Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. *Nature methods*, 17(2):175–183, 2020.

Dalla-Torre, H., Gonzalez, L., Mendoza-Revilla, J., Lopez Carranza, N., Grzywaczewski, A. H., Oteri, F., Dallago, C., Trop, E., de Almeida, B. P., Sirelkhatim, H., et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

Das, S. and Chakrabarti, S. Classification and prediction of protein–protein interaction interface using machine learning algorithm. *Scientific reports*, 11(1):1761, 2021.

Despond, E. A. and Dawson, J. F. Classifying cardiac actin mutations associated with hypertrophic cardiomyopathy. *Frontiers in Physiology*, 9:405, 2018.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pp. 2021–10, 2021.

Fang, X., Liu, L., Lei, J., He, D., Zhang, S., Zhou, J., Wang, F., Wu, H., and Wang, H. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

Feng, S., Ni, Y., Lan, Y., Ma, Z.-M., and Ma, W.-Y. Fractional denoising for 3d molecular pre-training. In *International Conference on Machine Learning*, pp. 9938–9961. PMLR, 2023.

Fey, M. and Lenssen, J. E. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*, 2019.

Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M., and Correia, B. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

Gainza, P., Wehrle, S., Van Hall-Beauvais, A., Marchand, A., Scheck, A., Harteveld, Z., Buckley, S., Ni, D., Tan, S., Sverrisson, F., et al. De novo design of protein interactions with learned surface fingerprints. *Nature*, pp. 1–9, 2023.

Gane, A., Bileschi, M. L., Dohan, D., Speretta, E., Héliou, A., Meng-Papaxanthos, L., Zellner, H., Brevdo, E., Parikh, A., Martin, M. J., Orchard, S., UniProt Collaborators, and Colwell, L. J. Protnlm: Model-based natural language protein annotation. 2022. URL https://storage.googleapis.com/brain-genomics-public/research/proteins/protnlm/uniprot_2022_04/protnlm_preprint_draft.pdf.

Geiger, M. and Smidt, T. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

Godwin, J., Schaarschmidt, M., Gaunt, A., Sanchez-Gonzalez, A., Rubanova, Y., Veličković, P., Kirkpatrick, J., and Battaglia, P. Simple gnn regularisation for 3d molecular property prediction & beyond. *arXiv preprint arXiv:2106.07971*, 2021.

Godwin, J., Schaarschmidt, M., Gaunt, A. L., Sanchez-Gonzalez, A., Rubanova, Y., Veličković, P., Kirkpatrick, J., and Battaglia, P. Simple GNN regularisation for 3d molecular property prediction and beyond. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=1wVvweK3oIb.

Google DeepMind AlphaFold Team and Isomorphic Labs Team. Performance and structural coverage of the latest, in-development alphafold model, oct 2023. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/a-glimpse-of-the-next-generation-of-alphafold/alphafold_latest_oct2023.pdf. Accessed: 2023-04-16.

Groom, C. R., Bruno, I. J., Lightfoot, M. P., and Ward, S. C. The cambridge structural database. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*, 72(2):171–179, 2016.

Hagberg, A. A., Schult, D. A., and Swart, P. J. Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J. (eds.), *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pp. 11–15, Pasadena, CA, USA, August 2008.

Harel, A., Bromberg, Y., Falkowski, P. G., and Bhattacharya, D. Evolutionary history of redox metal-binding domains across the tree of life. *Proceedings of the National Academy of Sciences*, 111(19):7042–7047, 2014.

Hekkelman, M. L., de Vries, I., Joosten, R. P., and Perrakis, A. Alphafill: enriching alphafold models with ligands and cofactors. *Nature Methods*, 20(2):205–213, 2023.

Hermosilla, P., Schäfer, M., Lang, M., Fackelmann, G., Vázquez, P. P., Kozlíková, B., Krone, M., Ritschel, T., and Ropinski, T. Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. *International Conference on Learning Representations*, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

Jiang, J., Wakimoto, H., Seidman, J., and Seidman, C. E. Allele-specific silencing of mutant myh6 transcripts in mice suppresses hypertrophic cardiomyopathy. *Science*, 342(6154):111–114, 2013.

Jiménez, J., Doerr, S., Martínez-Rosell, G., Rose, A. S., and De Fabritiis, G. Deepsite: protein-binding site predictor using 3d-convolutional neural networks. *Bioinformatics*, 33(19):3036–3042, 2017.

Jin, W., Chen, X., Vetticaden, A., Sarzikova, S., Raychowdhury, R., Uhler, C., and Hacohen, N. Dsmbind: Se (3) denoising score matching for unsupervised binding energy prediction and nanobody design. *bioRxiv*, pp. 2023–12, 2023.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Kandel, J., Tayara, H., and Chong, K. T. Puresnet: prediction of protein-ligand binding sites using deep residual neural network. *Journal of cheminformatics*, 13(1):1–14, 2021.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Klicpera, J., Becker, F., and Günnemann, S. GemNet: Universal directional graph neural networks for molecules. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Kong, X., Huang, W., and Liu, Y. Generalist equivariant transformer towards 3d molecular interaction learning. *arXiv preprint arXiv:2306.01474*, 2023.

Krapp, L. F., Abriata, L. A., Cortés Rodriguez, F., and Dal Peraro, M. Pesto: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nature Communications*, 14(1):2175, 2023.

Kratz, A., Kim, M., Kelly, M. R., Zheng, F., Koczor, C. A., Li, J., Ono, K., Qin, Y., Churas, C., Chen, J., et al. A multi-scale map of protein assemblies in the dna damage response. *Cell Systems*, 14(6):447–463, 2023.

Krishna, R., Wang, J., Ahern, W., Sturmfels, P., Venkatesh, P., Kalvet, I., Lee, G. R., Morey-Burrows, F. S., Anishchenko, I., Humphreys, I. R., et al. Generalized biomolecular modeling and design with rosettafold all-atom. *Science*, pp. eadl2528, 2024.

Kulkarni, P. and Uversky, V. N. Intrinsically disordered proteins: the dark horse of the dark proteome. *Proteomics*, 18(21-22):1800061, 2018.

Lam, J. H., Li, Y., Zhu, L., Umarov, R., Jiang, H., Héliou, A., Sheong, F. K., Liu, T., Long, Y., Li, Y., et al. A deep learning framework to predict binding preference of rna constituents on protein surface. *Nature communications*, 10(1):4941, 2019.

Leal, S. S., Botelho, H. M., and Gomes, C. M. Metal ions as modulators of protein conformation and misfolding in neurodegeneration. *Coordination Chemistry Reviews*, 256(19-20):2253–2270, 2012.

Lei, Y., Li, S., Liu, Z., Wan, F., Tian, T., Li, S., Zhao, D., and Zeng, J. A deep-learning framework for multi-level peptide–protein interaction prediction. *Nature communications*, 12(1):5465, 2021.

Levitt, M. Nature of the protein universe. *Proceedings of the National Academy of Sciences*, 106(27):11079–11084, 2009.

Li, L., Jamieson, K., Rostamizadeh, A., Gonina, E., Ben-Tzur, J., Hardt, M., Recht, B., and Talwalkar, A. A system for massively parallel hyperparameter tuning. *Proceedings of Machine Learning and Systems*, 2:230–246, 2020.

Li, S., Zhou, J., Xu, T., Huang, L., Wang, F., Xiong, H., Huang, W., Dou, D., and Xiong, H. Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity. KDD '21, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3447548.3467311.

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., and Stoica, I. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*, 2018.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H., and Tang, J. Pre-training molecular graph representation with 3D geometry. *International Conference on Learning Representations*, 2021a.

Liu, Y., Wang, L., Liu, M., Zhang, X., Oztekin, B., and Ji, S. Spherical message passing for 3d graph networks. *arXiv preprint arXiv:2102.05013*, 2021b.

Luo, S., Chen, T., Xu, Y., Zheng, S., Liu, T.-Y., Wang, L., and He, D. One transformer can understand both 2d & 3d molecular data. In *The Eleventh International Conference on Learning Representations*, 2022.

Matsuda, T., Jeong, J. I., Ikeda, S., Yamamoto, T., Gao, S., Babu, G. J., Zhai, P., and Del Re, D. P. H-ras isoform mediates protection against pressure overload–induced cardiac dysfunction in part through activation of akt. *Circulation: Heart Failure*, 10(2):e003658, 2017.

McNally, E. M. β-myosin heavy chain gene mutations in familial hypertrophic cardiomyopathy: The usual suspect? *Circulation Research*, 90(3):246–248, February 2002. doi: 10.1161/res.90.3.246. URL https://doi.org/10.1161/res.90.3.246. Originally Published 22 February 2002, Free Access.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29287–29303. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f51338d736f95dd42427296047067694-Paper.pdf.

Meller, A., Ward, M. D., Borowsky, J. H., Lotthammer, J. M., Kshirsagar, M., Oviedo, F., Ferres, J. L., and Bowman, G. Predicting the locations of cryptic pockets from single protein structures using the pocketminer graph neural network. *Nature Communications*, 2023.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., and Barabási, A.-L. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601, 2015.

Meng, Z. and Xia, K. Persistent spectral–based machine learning (perspect ml) for protein-ligand binding affinity prediction. *Science advances*, 7(19):eabc5329, 2021.

Moesser, M. A., Klein, D., Boyles, F., Deane, C. M., Baxter, A., and Morris, G. M. Protein-ligand interaction graphs: Learning from ligand-shaped 3d interaction graphs to improve binding affinity prediction. *bioRxiv*, pp. 2022–03, 2022.

Moon, S., Zhung, W., Yang, S., Lim, J., and Kim, W. Y. Pignet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chemical Science*, 13(13):3661–3673, 2022.

Musaelian, A., Batzner, S., Johansson, A., Sun, L., Owen, C. J., Kornbluth, M., and Kozinsky, B. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579, 2023.

Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Gonzalez-Uriarte, A., Malangone, C., Miranda, A., Fumis, L., Carvalho-Silva, D., Spitzer, M., et al. Open targets platform: supporting systematic drug–target identification and prioritisation. *Nucleic acids research*, 49 (D1):D1302–D1310, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., Signal, B., Gloss, B. S., Hammang, C. J., Rost, B., et al. Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences*, 112 (52):15898–15903, 2015.

Renaud, N., Geng, C., Georgievska, S., Ambrosetti, F., Ridder, L., Marzella, D. F., Réau, M. F., Bonvin, A. M., and Xue, L. C. Deeprank: a deep learning framework for data mining 3d protein-protein interfaces. *Nature communications*, 12(1):7068, 2021.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Rube, H. T., Rastogi, C., Feng, S., Kribelbauer, J. F., Li, A., Becerra, B., Melo, L. A., Do, B. V., Li, X., Adam, H. H., et al. Prediction of protein–ligand binding affinity from sequencing data with interpretable machine learning. *Nature Biotechnology*, 40(10):1520–1527, 2022.

Saliba, A.-E., Vonkova, I., and Gavin, A.-C. The systematic analysis of protein–lipid interactions comes of age. *Nature Reviews Molecular Cell Biology*, 16(12):753–761, 2015.

Sawicki, K. T., Chang, H.-C., and Ardehali, H. Role of heme in cardiovascular physiology and disease. *Journal of the American Heart Association*, 4(1):e001138, 2015.

Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A., and Müller, K.-R. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

Shan, L., Dauvilliers, Y., and Siegel, J. M. Interactions of the histamine and hypocretin systems in cns disorders. *Nature reviews Neurology*, 11(7):401–413, 2015.

Steinegger, M. and Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Sun, L., Xu, K., Huang, W., Yang, Y. T., Li, P., Tang, L., Xiong, T., and Zhang, Q. C. Predicting dynamic cellular protein–rna interactions by deep learning using in vivo rna structures. *Cell research*, 31(5):495–516, 2021.

Sverrisson, F., Feydy, J., Correia, B. E., and Bronstein, M. M. Fast end-to-end learning on protein surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15272–15281, 2021.

Tateishi-Karimata, H. and Sugimoto, N. Roles of non-canonical structures of nucleic acids in cancer and neurodegenerative diseases. *Nucleic Acids Research*, 49(14): 7839–7855, 2021.

Townshend, R. J., Eismann, S., Watkins, A. M., Rangan, R., Karelina, M., Das, R., and Dror, R. O. Geometric deep learning of rna structure. *Science*, 373(6558):1047–1051, 2021. doi: 10.1126/science.abe5650. URL https://www.science.org/doi/abs/10.1126/science.abe5650.

Tsaban, T., Varga, J. K., Avraham, O., Ben-Aharon, Z., Khramushin, A., and Schueler-Furman, O. Harnessing protein folding neural networks for peptide–protein docking. *Nature communications*, 13(1):176, 2022.

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., et al. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic acids research*, 52(D1):D368–D375, 2024.

Wang, L., Liu, H., Liu, Y., Kurtin, J., and Ji, S. Learning protein representations via complete 3d graph networks. *arXiv preprint arXiv:2207.12600*, 2022a.

Wang, Y., Wang, J., Cao, Z., and Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022b.

Wei, H., Wang, W., Peng, Z., and Yang, J. Q-biolip: A comprehensive resource for quaternary structure-based protein–ligand interactions. *bioRxiv*, pp. 2023–06, 2023.

Wei, J., Chen, S., Zong, L., Gao, X., and Li, Y. Protein–rna interaction prediction with deep learning: structure matters. *Briefings in bioinformatics*, 23(1):bbab540, 2022.

Wollschläger, T., Kemper, N., Hetzel, L., Sommer, J., and Günnemann, S. Expressivity and generalization: Fragment-biases for molecular gnns. *arXiv preprint arXiv:2406.08210*, 2024.

Xia, Y., Xia, C.-Q., Pan, X., and Shen, H.-B. Graphbind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. *Nucleic acids research*, 49(9): e51–e51, 2021.

Yan, J., Ye, Z., Yang, Z., Lu, C., Zhang, S., Liu, Q., and Qiu, J. Multi-task bioassay pre-training for protein-ligand binding affinity prediction. *arXiv preprint arXiv:2306.04886*, 2023.

Yang, J., Roy, A., and Zhang, Y. Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, 41(D1):D1096–D1103, 2012.

Yen, H.-Y., Hoi, K. K., Liko, I., Hedger, G., Horrell, M. R., Song, W., Wu, D., Heine, P., Warne, T., Lee, Y., Carpenter, B., Plückthun, A., Tate, C. G., Sansom, M. S. P., and Robinson, C. V. Pip2 stabilises active states of gpcrs and enhances the selectivity of g-protein coupling. *Nature*, 559(7714):423–427, Jul 2018a. doi: 10.1038/s41586-018-0325-6. URL https://www.nature.com/articles/s41586-018-0325-6. Author

manuscript; available in PMC: 2019 Jan 11. Published in final edited form as: Nature. 2018 Jul 11.

Yen, H.-Y., Hoi, K. K., Liko, I., Hedger, G., Horrell, M. R., Song, W., Wu, D., Heine, P., Warne, T., Lee, Y., et al. Ptdins (4, 5) p2 stabilizes active states of gpcrs and enhances selectivity of g-protein coupling. *Nature*, 559 (7714):423–427, 2018b.

Zaidi, S., Schaarschmidt, M., Martens, J., Kim, H., Teh, Y. W., Sanchez-Gonzalez, A., Battaglia, P., Pascanu, R., and Godwin, J. Pre-training via denoising for molecular property prediction. In *The Eleventh International Conference on Learning Representations*, 2022.

Zaidi, S., Schaarschmidt, M., Martens, J., Kim, H., Teh, Y. W., Sanchez-Gonzalez, A., Battaglia, P., Pascanu, R., and Godwin, J. Pre-training via denoising for molecular property prediction. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tYIMtogyee.

Zheng, F., Kelly, M. R., Ramms, D. J., Heintschel, M. L., Tao, K., Tutuncuoglu, B., Lee, J. J., Ono, K., Foussard, H., Chen, M., et al. Interpretation of cancer mutations using a multiscale map of protein systems. *Science*, 374 (6563):eabf3067, 2021.

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., and Ke, G. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

A. Construction of hierarchical graphs of interacting molecules

Given the atomic structure of two molecules interacting, an atom-level graph is then constructed. Each atom in the complex maps to an atom node in the graph with the features: element and 3D coordinates of the atom. Intramolecular atom edges are defined for each atom to the k nearest atoms in the same molecule. Intermolecular atom edges are defined for each atom to the k nearest atoms in the other molecule. In total, there are 118 atom types based on the elements of the periodic table.

Atom nodes are connected to the next level of nodes, block nodes. Block nodes have the features: block type and 3D coordinates of the block given by the mean of the atomic coordinates of atoms in the block. Each atom is connected to one block node. For proteins, peptides, DNA, and RNA, we define the atoms that belong to a given block by the amino acid and nucleotide residues. For small molecule ligands, blocks are defined by a vocabulary of 290 common chemical motifs. Atoms of sections of the molecule that cannot be fragmented into these motifs become blocks comprised of one atom. We use the vocabulary and fragmentation of the molecule to blocks from (Kong et al., 2023). Intramolecular block edges are defined for each atom to the k nearest blocks in the same molecule. Intramolecular block edges are defined for each atom to the k nearest blocks in the other molecule. In total, there are the following block types: 20 for canonical amino acids, 4 for DNA nucleotides, 4 for RNA nucleotides, 290 for small molecule fragments, and 118 for elemental blocks.

In addition, there are three special block types: mask, unknown, and global. The mask node is applied at pretraining for masked identity prediction of blocks. Unknown nodes are used for nodes that do not fall into the defined vocabulary, such as non-canonical amino acids and nucleotides. There are also two atom global-type nodes at the atom and block level. The two global nodes are connected to all nodes in each molecule at their respective level.

B. All-atom graph neural network

ATOMICA uses a SE(3)-equivariant 3D message passing network on graphs of molecular complexes to learn representations that are informative of the intermolecular interactions between molecules.

B.1. Atom-level representation learning

Here we outline the SE(3)-equivariant 3D message passing network for ATOMICA on the nodes of the graph G^i. Several rotational equivariant neural networks have been introduced for modeling molecules (Schütt et al., 2018; Klicpera et al., 2021; Liu et al., 2021b; Batzner et al., 2022). We build on the E(3)-equivariant neural network layers presented by Tensor-Field Networks implemented in e3nn (Geiger & Smidt, 2022) and DiffDock (Corso et al., 2023). Message passing for the intermolecular edges and intramolecular edges is done separately, but the message passing framework for the two edge types is the same.

The feature vector of atom ($\mathbf{h}_a^{\text{atom}}$) node a in G^i is a geometric object comprised of a direct sum of irreducible representations of the O(3) symmetry group. The feature vectors $\mathbf{h}_{a,(\lambda,p)}^{\text{atom}}$ are indexed with λ, p, where $\lambda = 0, 1, 2, \ldots$ is a non-negative integer denoting the rotation order and $p \in \{\text{o}, \text{e}\}$ indicates odd or even parity, which together index the irreducible representations (irreps) of O(3). In our model, we set $\lambda_{\max} = 1$ for $\mathbf{h}_a^{\text{atom}}$, and we denote the number of scalar (0e) and pseudoscalar (0o) irrep features in $\mathbf{h}_a^{\text{atom}}$ with ns, and the number of vector (1o) and pseudovector (1e) irrep features in $\mathbf{h}_a^{\text{atom}}$ with nv.

The atom-type of node a, determined by the element of the atom, is embedded with a normal distribution and trainable weights as a scalar $\text{ns} \times 0\text{e}$. There are L_{GNN} layers of message passing between atom nodes. At each layer l, the node updates for node a in the graph of interaction complex G^i are given by:

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

$$\text{with } \psi_{ab} = \Psi\left(\mathbf{e}_{ab}, \mathbf{t}_{ab}, \mathbf{h}_{a,(0e)}^{\text{atom}}, \mathbf{h}_{b,(0e)}^{\text{atom}}\right). \tag{2}$$

After each layer l of message passing, $\mathbf{h}_a^{\text{atom}}$ is filtered down to irreps with $\lambda_{\max} = 2$. After L layers the $\mathbf{h}_a^{\text{atom}}$ embedding is projected with a 2-layer MLP to a d_{node}-dimension vector.

B.2. Block-level representation learning

The feature vector of block ($\mathbf{h}_b^{\text{block}}$) node b in G^i is also a geometric object defined in the same way as ($\mathbf{h}_a^{\text{atom}}$). We initialize block nodes using a scalar, $\text{ns} \times 0\text{e}$, trainable embedding of block types.

Let d_{node} be the dimension of $\mathbf{h}_b^{\text{block}}$ and n_{heads} be the number of attention heads. We define $d_h = d_{\text{node}}/n_{\text{heads}}$ as the dimension per head. The multi-head cross-attention operation can be expressed as:

$$\mathbf{h}_b^{\text{block}} \leftarrow \mathbf{h}_b^{\text{block}} + \text{MultiHead}(\mathbf{h}_b^{\text{block}}, \{\mathbf{h}_a^{\text{atom}}\}_{a \in A_b}) \tag{3}$$

where A_b is the set of atoms in block b, and MultiHead is defined as:

$$\text{MultiHead}(\mathbf{h}_b^{\text{block}}, \{\mathbf{h}_a^{\text{atom}}\}_{a \in A_b}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_{n_{\text{heads}}})\mathbf{W}_O \tag{4}$$

and each head computed as:

$$\text{head}_i = \sum_{a \in A_b} \alpha_{ba} \mathbf{v}_a^{\text{atom},(i)} \text{ with } \alpha_{ba} = \frac{\exp\left(\mathbf{q}_b^{\text{block},(i)} \cdot \mathbf{k}_a^{\text{atom},(i)} / \sqrt{d_h}\right)}{\sum_{v \in A_b} \exp\left(\mathbf{q}_b^{\text{block},(i)} \cdot \mathbf{k}_v^{\text{atom},(i)} / \sqrt{d_h}\right)} \tag{5}$$

where $\mathbf{q}_b^{\text{block},(i)} = \mathbf{h}_b^{\text{block}}\mathbf{W}_Q^{(i)}$, $\mathbf{k}_a^{\text{atom},(i)} = \mathbf{h}_a^{\text{atom}}\mathbf{W}_K^{(i)}$, $\mathbf{v}_a^{\text{atom},(i)} = \mathbf{h}_a^{\text{atom}}\mathbf{W}_V^{(i)}$, and $\mathbf{W}_Q^{(i)}, \mathbf{W}_K^{(i)}, \mathbf{W}_V^{(i)} \in \mathbb{R}^{d_{\text{node}} \times d_h}$ and $\mathbf{W}_O \in \mathbb{R}^{d_{\text{node}} \times d_{\text{node}}}$. Message passing between the block nodes follows the same architecture as the atom nodes described in equation 1 with separate model parameters.

B.3. Graph-level representation learning

To pool $\mathbf{h}_b^{\text{block}} \in \mathbb{R}^d$ for $b \in G^i$ for a graph-level representation $\mathbf{h}_i^{\text{graph}} \in \mathbb{R}^d$, we use multi-head self-attention for L_{pool} layers and sum the output $\mathbf{h}_b^{\text{block}}$ for all $b \in G^i$ for $\mathbf{h}_i^{\text{graph}}$.

C. Self-supervised learning on interaction complexes

C.1. Geometric Denoising

Node-level denoising as an objective function has been useful for pretraining on 3D coordinate molecular datasets from DFT generated molecules to prevent over-smoothing of GNNs (Godwin et al., 2021), and it has proven that it is related to learning a force field of per-atom forces (Zaidi et al., 2022; Feng et al., 2023). In addition, denoising is linked to score-matching which has also been popular in training generative models (Ho et al., 2020; Corso et al., 2023) as well as unsupervised binding affinity prediction (Jin et al., 2023). Thus, this motivates the application of denoising as an objective for self-supervised training.

To predict the rotation score $\mathbf{s}_{\boldsymbol{\omega}} \in \mathbb{R}^3$ and the translation score $\mathbf{s}_{\mathbf{t}} \in \mathbb{R}^3$ from $\widetilde{G}^i$, the node representations at the atom and block level are convolved with the center of the graph using a tensor field network (Corso et al., 2023):

$$\mathbf{s} \leftarrow \text{LN}\left(\frac{1}{|\mathcal{A}'|} \sum_{a \in \mathcal{A}'} Y\left(\widehat{r}_{ca}\right) \otimes_{\phi_{ca}} \mathbf{h}_a^{\text{atom}}\right) \text{ with } \phi_{ca} = \Phi\left(\mathbf{e}_{ca}, \mathbf{h}_{a,(0e)}^{\text{atom}}\right), \tag{6}$$

where node $a \in \mathcal{A}'$ are the atom nodes in the perturbed molecule and c is the center of the perturbed molecule. This is a weighted tensor product, with the weights given by a 2-layer MLP, Φ, which takes as input the Gaussian smearing d_{edge}-embedding of the Euclidean distance between coordinates of the center c and node a, and the scalar component of $\mathbf{h}_a^{\text{atom}}$.

Finally, the rotation score is given by the pseudovector irrep component $\mathbf{s}_{\boldsymbol{\omega}} = \Gamma_{\boldsymbol{\omega}}(\mathbf{h}_i^{\text{graph}}) * \mathbf{s}_{(1e)}$ and the translation score is given by the vector irrep component $\mathbf{s}_{\mathbf{t}} = \Gamma_{\mathbf{t}}(\mathbf{h}_i^{\text{graph}}) * \mathbf{s}_{(1o)}$, where $\Gamma_{\boldsymbol{\omega}}$ and $\Gamma_{\mathbf{t}}$ are 2-layer MLPs that project the graph representation of G^i to a single scalar.

To predict the torsion score $\mathbf{s}_{\boldsymbol{\theta}} \in \mathbb{R}^m$ the atom nodes are convolved with the center of the rotatable bonds connecting atoms a_{z0}, a_{z1}. Let z denote the center of one of the rotatable bonds. We connect $\mathcal{N}_z = \{a \,|\, a \in \mathcal{A}', ||\widehat{r}_{za}|| < 5\mathring{A}\}$ which is all

atoms in the perturbed molecule within 5 Å to the center of the bond.

$$\mathbf{h}_z = \frac{1}{|\mathcal{N}_z|} \sum_{a \in \mathcal{N}_z} \left(Y^2(\hat{\mathbf{r}}_z) \otimes Y(\hat{\mathbf{r}}_{za}) \right) \otimes_{\pi_{za}} \mathbf{h}_a^{\text{atom}} \text{ with } \pi_{za} = \Pi^{(t)} \left(\mathbf{e}_{za}, \mathbf{h}_{a,(0e)}^{\text{atom}}, \mathbf{h}_{a_{z0},(0e)}^{\text{atom}} + \mathbf{h}_{a_{z1},(0e)}^{\text{atom}} \right), \tag{7}$$

The first tensor product is between the second order irreps of the unit direction vector along the two atoms a_{z0}, a_{z1} of the bond z, $Y^2(\hat{\mathbf{r}}_z)$, and the unit direction vector between the center of the bond and atom a, $Y(\hat{\mathbf{r}}_{za})$. This is followed by a weighted tensor product with the weights given by a 2-layer MLP, Π, which takes as input the Gaussian smearing d_{edge}-embedding of the Euclidean distance between coordinates of the bond center z and node a, the scalar component of $\mathbf{h}_a^{\text{atom}}$, and the sum of the scalar component of the two atoms in the bond $\mathbf{h}_{a_{z0}}, \mathbf{h}_{a_{z1}}$. Finally, we sum the scalar and pseudoscalar components of $\mathbf{h}_z$ and project it to a single scalar $\mathbf{s}_{\boldsymbol{\theta}_z}$ using a 2-layer MLP.

We calculate the loss components with:

$$l_\omega = ||\mathbf{s}_{\boldsymbol{\omega}} - \nabla_{\boldsymbol{\omega}} \log p(\boldsymbol{\omega})||^2 \tag{8}$$

$$l_t = ||\mathbf{s}_{\mathbf{t}} - \nabla_{\mathbf{t}} \log p(\mathbf{t})||^2 \tag{9}$$

$$l_\theta = \sum_z ||\mathbf{s}_{\boldsymbol{\theta}_z} - \nabla_{\boldsymbol{\theta}_z} \log p(\boldsymbol{\theta}_z)||^2 \tag{10}$$

where $\nabla_{\mathbf{t}} \log p(\mathbf{t}) = -\mathbf{t}/\sigma_t^2$. The values of $\nabla_{\mathbf{t}} \log p(\mathbf{t}), \nabla_{\boldsymbol{\theta}_z} \log p(\boldsymbol{\theta}_z)$ can be calculated by pre-computing a truncated infinite series following (Corso et al., 2023; Jin et al., 2023).

C.2. Masking Blocks

In addition to denoising, we also pretrain the model by masking out block identities and predicting the masked block identities. For each graph G^i, 10% of blocks are randomly sampled and their block identities are replaced with the special 'mask' block and we denote these blocks as $\mathcal{B}$. For a masked block $b \in \mathcal{B}$, the probability vector of the block identity is predicted with $\hat{\mathbf{y}}_b = \text{Softmax}(\Upsilon(\mathbf{h}_b^{\text{block}}))$, where Υ is a 2-layer MLP. We calculate the masked loss using a cross-entropy loss:

$$l_m = -\frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \mathbf{y}_b \cdot \log(\hat{\mathbf{y}}_b) \tag{11}$$

C.3. Loss Function

The pretraining loss is then calculated by a weighted sum of the above loss functions:

$$\mathcal{L} = \beta_\omega l_\omega + \beta_t l_t + \beta_\theta l_\theta + \beta_m l_m$$

.

D. Curation of pretraining dataset

D.1. Small molecule structures

We extract structures of small molecule interactions from the Cambridge Structural Database (CSD) v2023.2.0. The database was filtered for all CSD entries that satisfied the following criteria: organic, not polymeric, has 3D coordinates, no disorder, no errors, no metals, had only one SMILES string describing the crystal entry (in other words, each crystal is comprised of only one chemical compound), and molecules with 6-50 heavy atoms. CSD entries are unit cells of infinitely repeating crystal lattices. For our purposes of learning intermolecular interactions, we sampled many patches of intermolecular interactions to represent all examples of intermolecular interactions in a given unit cell. Given an entry of the CSD, we iterate through each unique conformer in the unit cell and extract all pairs of interactions with neighboring peripheral conformers that are within 4 Å to the central conformer using the CSD Python API. In total, there are 1,767,710 structures of molecular pairs from 375,941 CSD entries. Inspired by fingerprint-based similarity measures used in chemistry (Bajusz et al., 2015), we use a one-hot encoding of the molecular complex from a vocabulary of 290 common chemical motifs (Kong et al., 2023) and Manhattan distance between the embeddings to sample 1,000 molecular complexes and their 100 nearest neighbors, giving a total of 10,000 molecular complexes for validation and test splits respectively, that are distinct from the training set.

D.2. Biomolecular structures

We extract structures of interacting molecules from QBioLiP (June 2024), this includes structures of proteins interacting with ions, ligands, DNA, RNA, peptides, and proteins, and nucleic acids interacting with ions and ligands from the Protein Databank (PDB). For proteins, peptides, DNA, and RNA, we crop the complex to keep all residues within 8 Å to any atom, amino acid, or nucleic acid residue in the other molecule. In total, there are 124,541 protein-protein interaction complexes, 119,017 protein-small molecule interaction complexes, 74,514 pro-tein-ion interaction complexes, 8,475 protein-peptide interaction complexes, 5,185 nucleic acid-ligand interaction complexes, 3,511 protein-RNA interaction complexes, and 2,750 protein-DNA interaction complexes. For protein-ion, protein-small molecule, protein-peptide, and protein-protein molecular complexes, we cluster each modality with 30% protein sequence similarity using MMseqs2 with a coverage of 80%, sensitivity of 8, and cluster mode 1(Steinegger & Söding, 2017). For protein-protein complexes, we also ensure that for any two complexes in different clusters there is a maximum of 30% sequence similarity between all chains in the two complexes. For protein-RNA and protein-DNA complexes, we cluster by 30% protein sequence similarity and 30% nucleotide similarity using MMseqs2 with the same settings as above, this ensures that complexes in different clusters have a maximum of 30% protein sequence similarity and 30% nucleotide sequence similarity. For nucleic acid-ligand structures, we cluster based on 30% nucleotide sequence similarity. Finally, we split clusters into train, validation, and test splits using an 8:1:1 ratio.

E. Training details for ATOMICA

We pretrain ATOMICA on the training split of biomolecular structures and small molecule structures to generate embeddings of molecular complexes at the atom, block, and graph scale. To learn representations in a self-supervised manner, during training, we apply noise to the atomic coordinates and mask block identities of the input graphs of the molecular complex. At inference time, embeddings from the graphs are generated without noise or masked blocks.

E.1. Hyperparameter tuning

We employed a hyperparameter optimization strategy utilizing Ray Tune (Liaw et al., 2018) in conjunction with Optuna (Akiba et al., 2019) and the Asynchronous Successive Halving Algorithm (ASHA) scheduler (Li et al., 2020). The hyperparameter space we search on includes: the number of nearest neighbors to define edges to in the graph $k \in [4, \mathbf{8}, 16]$, dropout in the tensor field network $\in [\mathbf{0.00}, 0.01, 0.05, 0.10]$, edge dimension $d_{\text{edge}} \in [\mathbf{16}, 24, 32]$, node dimension $d_{\text{node}} \in [16, 24, \mathbf{32}]$, and the number of tensor field network layers $L \in [\mathbf{4}, 6, 8]$. The best hyperparameters are shown in bold and chosen based on the lowest validation loss when trained on a random 10% subsample of the training set. Then for determining the level of noise to apply to the interaction complexes, we conducted a second hyperparameter search on rotation $\sigma_\omega \in [\mathbf{0.25}, 0.5, 1]$, $\omega_{\text{max}} \in [0.25, \mathbf{0.5}, 1]$, translation $\sigma_t \in [0.5, \mathbf{1}, 1.5]$, and torsion $\sigma_\theta \in [0.25, \mathbf{0.5}, 1]$. The best hyperparameters are shown in bold and chosen based on the highest masked block identity prediction accuracy when trained on a random 10% subsample of the training set. For the loss function, we set $\beta_\omega = 1, \beta_t = 1, \beta_m = 0.1$, and the block identities are randomly masked at 10% probability. ATOMICA is trained on the full training set with the above hyperparameters, the learning rate cycles between 1e-4 and 1e-6 using Cosine Annealing Warm Restarts, with a cycle length of 400,000 steps, and the model is trained for 150 epochs.

E.2. Implementation

ATOMICA is implemented with PyTorch (Version 2.1.1) (Paszke et al., 2019) and PyTorch Geometric (Version 2.1.1) (Fey & Lenssen, 2019). Training runs were monitored with Weights and Biases (Biewald, 2020). Models are trained on 4 NVIDIA H100 Tensor Core GPUs in parallel.

E.3. Training ATOMICA on a single pair of interacting modalities

To demonstrate representations learned by ATOMICA are generalizable across multiple modalities, we train models with identical architecture and hyperparameters on only single pairs of interacting modalities (small molecules, protein-ion, protein-small molecule, protein-DNA, protein-RNA, protein-peptide, protein-protein, nucleic acid-small molecule). Using the same training set-up as ATOMICA, these models are trained on the same training data as ATOMICA but filtered for only one pair of interacting modalities. The models are trained for 150 epochs on 4 NVIDIA H100 Tensor Core GPUs in parallel. The model checkpoint with the lowest validation loss is then used for further finetuning on masked block identity

prediction on the same training data for 50 epochs with a learning rate of 1e-4. We also finetune ATOMICA for 50 epochs on block identity prediction for each pair of interacting modalities. To compare the quality of embeddings generated by ATOMICA and versions of it trained on single modalities, we evaluate the accuracy of masked block identity prediction on a test set. This test set was not seen by any of the models and has 30% sequence similarity and minimal small molecule fingerprint similarity to any training and validation data.

F. Interfaceome

F.1. Training ATOMICA-Interface

To support the embedding of protein binding interfaces, we finetune ATOMICA with structures of interfaces rather than complexes. For our finetuning dataset, we adapt the biomolecular structures from the training set. For each graph G^i we crop the graph to only the protein interface of one protein in the complex $G^{i\prime}$. Let $\mathbf{h}_i^{\text{graph}} = \mathcal{F}(G^i)$ where $\mathcal{F}$ is pretrained and frozen ATOMICA. Our goal is to train $\mathcal{G}$ initialized with $\mathcal{F}$ such that $\mathbf{h}_i^{\text{graph}\,\prime} = \mathcal{G}(G^{i\prime})$ for every intermolecular patch G^i such that $\mathbf{h}_i^{\text{graph}}$ and $\mathbf{h}_i^{\text{graph}\,\prime}$ are aligned. Then for a randomly sampled mini-batch of N_{batch} examples, the loss function is:

$$\mathcal{L}_{\text{interface}} = -\sum_{i=1}^{N_{\text{batch}}} \log \frac{\exp\left(\text{sim}\left(\mathbf{h}_i^{\text{graph}}, \mathbf{h}_i^{\text{graph}\,\prime}\right)/\tau\right)}{\sum_{j=1}^{N_{\text{batch}}} \mathbb{1}_{[j\neq i]} \exp\left(\text{sim}\left(\mathbf{h}_i^{\text{graph}}, \mathbf{h}_j^{\text{graph}\,\prime}\right)/\tau\right)} + \log \frac{\exp\left(\text{sim}\left(\mathbf{h}_i^{\text{graph}\,\prime}, \mathbf{h}_i^{\text{graph}}\right)/\tau\right)}{\sum_{j=1}^{N_{\text{batch}}} \mathbb{1}_{[j\neq i]} \exp\left(\text{sim}\left(\mathbf{h}_j^{\text{graph}}, \mathbf{h}_i^{\text{graph}\,\prime}\right)/\tau\right)} \tag{12}$$

where sim is cosine similarity and τ is the temperature factor. This contrastive loss is adapted from the normalized temperature-scaled cross-entropy loss (Chen et al., 2020). We finetune the model for 50 epochs with a cyclic learning rate ranging from 1e-3 to 1e-5 over 50000 steps. Three replicates of the model are trained. The models were finetuned on 4 NVIDIA H100 Tensor Core GPUs in parallel.

F.2. Detection of binding sites across the human proteome with PeSTo

We employ PeSTo (Version 4.1) (Krapp et al., 2023), which for a given protein structure, PeSTo predicts the probability of each amino acid as a binding site for an ion, ligand, nucleic acid, protein, and lipid binder. PeSTo is run across all human proteins from the AlphaFold Protein Structure Database (Varadi et al., 2024; Jumper et al., 2021). For each protein and binding modality, we extract binding sites as all amino acids with PeSTo confidence > 0.7 and AlphaFold2 pLDDT > 70 with at least 5 amino acids at the binding site to keep only high-confidence binding sites. This gives us a total of 6,458 protein-ion binding interfaces, 5,856 protein-ligand binding interfaces, 6,649 protein-nucleic acid binding sites, 6,766 protein-lipid binding sites, and 17,158 protein-protein binding interfaces.

F.3. Therapeutic targets dataset

We extract targets for diseases from Open Targets (2024-09) (Ochoa et al., 2021). Genes are associated with diseases using multiple lines of evidence (genetic association, somatic mutations, known drug, affected pathway) and we use the overall score, which is an aggregated sum of all evidence sources. For all diseases, we keep all targets with overall evidence scores > 0.5.

G. Dark proteome binding site characterization with ATOMICA-Ligand

We demonstrate versatility in ATOMICA and finetune the model for annotating ions and ligands to binding sites. The finetuned version of the model is applied to putative binding sites in the dark proteome.

G.1. Training ATOMICA-Ligand

The objective is to predict the probability of a specific ion or ligand binding to a given protein interface pocket. We frame this as a binary prediction task and finetune a separate model for each ion and small molecule. A predictive head is a L_{ligand}-layer MLP. For each ion and small molecule, we use RayTune with Optuna and ASHA to finetune ATOMICA-Ligand from

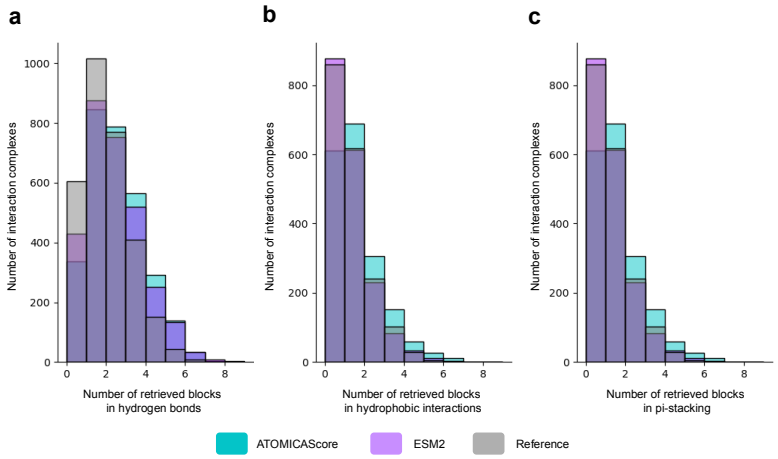

Figure S1. **Number of blocks involved in intermolecular bonds in the top 10 nominated blocks for protein-small molecule complexes in the pretraining test set.** We compare ATOMICASCORE, ESM-2 (3B), and a reference on the recovery of the following types of intermolecular bonds: **a** hydrogen bonds, **b** hydrophobic interactions, and **c** pi-stacking.

ATOMICA and find the optimal hyperparameters among $L_{\text{ligand}} \in [3, 4, 5]$, learning rate $\in [10^{-6}, 10^{-3}]$, non-linearity $\in$ [`relu, gelu, elu`], hidden dimension of MLP $\in [16, 32, 64]$, gradient clipping $\in$ [`None`, 1], and the number of nearest neighbors to define edges to in the graph $k \in [4, 6, 8]$. To address class imbalances in our dataset, we apply a weighted sampling strategy during training, where each protein pocket receives a sampling weight inversely proportional to the total count of its label class. For each ion and small molecule, we finetune ATOMICA-Ligand for 50 epochs on 1 NVIDIA H100 Tensor Core GPU. Three replicate models are trained for each ion and small molecule. For binary classification of binding sites, we set thresholds that maximize the F1 score, constraining these values to fall within the range of 0.05 to 0.95.

G.2. Dataset curation

Given an ion or small molecule, we separate all graphs in the pretraining set containing this ion bound to a protein. We cluster protein binders with a 30% protein sequence similarity cutoff, coverage of 80%, sensitivity of 8, and cluster mode 1 using MMseqs2 (Steinegger & Söding, 2017). The clusters are then divided into training, validation, and test sets in an 8:1:1 ratio. We set up this split for the following metal ions: Ca, Co, Cu, Fe, K, Mg, Mn, Na, Zn, and the following small molecules with these PDB chemical codes: ADP (adenosine diphosphate), ATP (adenosine triphosphate), CIT (citric acid), CLA (chlorophyll A), FAD (flavin adenine dinucleotide), GDP (guanosine diphosphate), GTP (guanosine triphosphate), HEC (heme C), HEM (heme B), NAD (nicotinamide adenine dinucleotide), NAP (NADP+, nicotinamide adenine dinucleotide phosphate, oxidized form), NDP (NADPH, nicotinamide adenine dinucleotide phosphate, reduced form).

G.3. Dark proteome annotation

The dark proteome is comprised of proteins that are dissimilar in sequence and structure from all currently annotated proteins. We use the clusters of the dark proteome from FoldSeek cluster on the AlphaFold Protein Structure Database (Barrio-Hernandez et al., 2023). We limit our analysis to the 33,482 clusters with an average pLDDT > 90. For each cluster, we take the representative protein and run PeSTo on the protein structure to predict ion and small molecule binding sites. We keep residues with PeSTo confidence > 0.8 as the putative binding site, with a minimum of 5 residues required. In total, we extract 2,851 ion binding proteins and 969 small molecule binding proteins from the 33,482 representative proteins. Given these binding interfaces, we run ATOMICA-Ligand for all finetuned ion and small molecules to annotate chemical identities to the binding sites. We evaluated the quality of our predicted protein-ligand complexes by folding them with AlphaFold3 and evaluating their ipTM scores. For comparison, we established a reference baseline using randomly sampled proteins from the dark proteome with predicted ion and small molecule binding capabilities. These reference proteins were selected and paired with ligands to match both the number and identity of annotated ligands in our predicted complexes. For sequence-based annotation we run the Google Colab notebook with ProtNLM (Gane et al., 2022).