ATOMICA: Learning Universal Representations of Intermolecular Interactions

Ada Fang[1,4], Zaixi Zhang[2], Andrew Zhou[2,3], and Marinka Zitnik[2,4,5,6,‡]

[1]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA
[2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[3]Program in Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, USA
[4]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, MA, USA
[5]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[6]Harvard Data Science Initiative, Cambridge, MA, USA

‡Corresponding author: marinka@hms.harvard.edu

Molecular interactions underlie nearly all biological processes, but most machine learning models treat molecules in isolation or specialize in a single type of interaction, such as protein-ligand or protein-protein binding. This siloed approach prevents generalization across biomolecular classes and limits the ability to model interaction interfaces systematically. 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 generalizes across molecular classes and recovers shared physicochemical features without supervision. Its latent space captures compositional and chemical similarities across interaction types and follows scaling laws that improve representation quality with increasing biomolecular data modalities. We apply ATOMICA to construct five modality-specific interfaceome networks, termed ATOMICANETs, which connect proteins based on interaction similarity with ions, small molecules, nucleic acids, lipids, and proteins. These networks identify disease pathways across 27 conditions and predict disease-associated proteins in autoimmune neuropathies and lymphoma. Finally, we use ATOMICA to annotate the dark proteome—proteins lacking known structure or function—by predicting 2,646 previously uncharacterized ligand-binding sites. These include putative zinc finger motifs and transmembrane cytochrome subunits, demonstrating that ATOMICA enables systematic annotation of molecular interactions across the proteome.

Main

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 (Fig. 1a) or provide limited modeling of molecular interactions, typically restricted to a specific type of interaction, such as protein-ligand and protein-protein interactions. These methods rely on separate architectures for different molecular classes, preventing cross-modality knowledge transfer and limiting the generalizability of learned representations. Methods for protein docking [1, 2] and protein structure prediction [3–5] have demonstrated success in specific domains. Machine learning has used atomic structures to predict protein folding from sequence [6–9], ligand binding [10–12], and de novo binder design [4, 13–15]. However, these methods remain specialized for individual molecular types. They focus on specific biomolecular modalities and do not generalize across the full diversity of molecular interactions. Current generative models, including AlphaFold [16] and RosettaFold [17], 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 [17, 18]. 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.

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 [3, 19–22], whereas small molecules require atomic-scale modeling due to their lack of inherent sequential structure [23–28]. As a result, molecular interaction models are specialized, with distinct architectures designed for protein-ligand binding affinity [29–33], binding site prediction [34–37], protein-peptide interactions [38–40], protein-protein interactions [41–46], and protein-RNA recognition [47–52]. This siloed approach prevents knowledge transfer across molecular classes even though interactions between proteins, nucleic acids, small molecules, and ions obey shared physicochemical principles. Structure-based generative models have demonstrated the feasibility of learning across all biomolecular modalities present in the Protein Data Bank [16, 17]. 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. A universal representation learning framework that integrates structural and chemical information across proteins, nucleic acids, small molecules, and ions remains an open challenge.

We introduce ATOMICA, an all-atom geometric 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 (Fig.1b). These include 1,747,710 small-molecule interaction complexes from the Cambridge Structural Database (CSD) [53] and 290,262 biomolecular complexes from Q-BioLiP and the Protein Data Bank (PDB) [54–56]. Interaction complexes across modalities share fundamental chemical forces—such as hydrogen bonding, van der Waals interactions, and electrostatics—making them a suitable foundation for universal representation learning. 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.

ATOMICA's learned representations capture critical residues at interaction interfaces, reflect chemically meaningful patterns, and exhibit semantic relationships. We apply ATOMICA on the human interfaceome to construct ATOMICANETs from putative interactions between 5,856 protein-small molecule, 6,458 protein-ion, 6,649 protein-nucleic acid, 6,766 protein-lipid, and 17,158 protein-protein interactions [57, 58]. Analysis of the interfaceome reveals that proteins with similar ATOMICA-derived interaction profiles often participate in shared disease pathways. We observe 27 disease-relevant pathways in ion, lipid, and small molecule networks, and associate high-confidence targets for autoimmune neuropathies and lymphoma. Moving beyond annotated proteins, we apply ATOMICA to the dark proteome—regions of the proteome lacking functional labels [59–61]. Finetuning ATOMICA enables the annotation of 2,646 binding sites with putative ions and cofactors, revealing functions in ancient and uncharacterized protein families.

Results

Universal modeling of intermolecular interactions at the atomic scale.

We model the interactions between molecules, which is contrary to prior work focused on modeling individual molecules or protein surfaces (Fig. 1a). 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. We assembled a dataset of pairs of interacting molecular entities from the Cambridge Structural Database (CSD) v2022.3.0 [53] and Q-BioLiP [54, 55] which includes all biologically relevant intermolecular interactions across all modalities available in the Protein Data Bank. We leverage the CSD for structures of interacting small molecules; the CSD was filtered for entries that satisfied the following criteria: organic, not polymeric, has 3D coordinates, no disorder, no errors, no metals, and was comprised of one compound. For each entry, we extract unique pairs of conformers from the unit cell where there are at least 5 atoms on a molecule within 4 Å of the other molecule. This results in 1,767,710 interacting pairs of molecules. We also extract all interfaces of interactions 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, this results in 337,993 interaction complexes. The interaction interface between two entities is defined by atoms within an 8 Å distance to the other molecule, to capture atoms in intermolecular bonds and surrounding molecular context. For a protein-small molecule, protein-ion, and nucleic acid-small molecule interaction complexes, we

model the binding site of the protein or nucleic acid and the whole ligand; for protein-nucleic acid and protein-protein interaction complexes, we model the amino acids and nucleic acids at the interface. In Fig. 1b, we show examples of the interaction interface for various complexes and their relative size.

We represent each interaction complex using a hierarchical graph that models both atomic-level details and higher-order chemical structure (Fig. 1c). 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 [62–64]. For example, the amino acid, phenylalanine, in a protein or a guanine nucleotide in RNA forms a block composed of its constituent atoms. This hierarchical design captures both local atomic interactions and broader structural organization and has theoretically higher expressive power than purely atom-level graphs [65]. 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.

ATOMICA geometric deep learning model.

ATOMICA is a self-supervised geometric graph neural network that generates multi-scale embeddings at the atom, block, and graph level from the structure of interacting two molecules (Fig. 1c). 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) (Fig. 1b). ATOMICA can also be adapted for other tasks by adding a task adaptor head and finetuning ATOMICA and the head for specific predictive tasks such as binding site annotation. Embeddings are generated with ATOMICA through message passing between nodes via the intermolecular and intramolecular edges. We use SE(3)-equivariant tensor field networks for message passing (Online Methods 2), which have been used to predict interatomic potentials [66, 67], molecular coupling [10], and scoring RNA structure [68]. Message passing is first done at the atom-level 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.

To learn high-quality representations, we employ a denoising strategy, which is effective as a pretraining objective to learn representations of 3D conformations of single molecules for property prediction [69–72] and unsupervised binding affinity prediction [73]. 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 (Fig. 1e). The model denoising output is optimized to minimize the distance to the score function of the global translation, global rotation, and torsion noise distributions [10, 73]. By denoising one molecular interface with respect to the other, this approach aims to capture the structural and geometric patterns of intermolecular interaction.

ATOMICA captures multi-scale relationships and composition of interacting molecules.

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. 1f). 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). The mean node embeddings of all nodes of each atomic element in the pretraining dataset are projected into 2D with principal component analysis (PCA), and we observe the organization of the latent space of the mean node embeddings according to periodic table location (Fig. 2a). Additionally, we see organization according to the chemical properties of the amino acids (Fig. 2b) 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. The ATOMICA latent space is complementary to existing molecular embedding models for representing protein-ligand interactions (ESM-2 3B [19], Spearman correlation of -0.144, p-value < 0.001), (Transformer-M [74], Spearman correlation of 0.107, p-value < 0.001).

Molecular structures are compositional in nature, with protein domains responsible for distinct functions [75] and the assembly of such domains into multi-domain proteins [76]. We examine if the graph embeddings exhibit compositional behavior by drawing inspiration from natural language processing. The latent space of word representations in word2vec exhibit similarity beyond simple syntactic regularities, that is, algebraic relationships such as $vector(\text{“}King\text{”}) - vector(\text{“}Man\text{”}) + vector(\text{“}Woman\text{”})$ results in an embedding closest to $vector(\text{“}Queen\text{”})$ exist [77, 78]. This compositional approach has also been shown in embeddings of genes with

their functions [79]. In intermolecular interactions, we examine the composition of protein-small molecule interactions. For a small molecule X that binds to different proteins A and B, we examine whether our model's latent space supports the relationship: ATOMICA Semantic Vector $:= vector(ProteinA, SmallMoleculeX) - vector(ProteinA) + vector(ProteinB) \simeq vector(ProteinB, SmallMoleculeX)$ (Fig. 2c). To test this property, we analyze previously unseen protein-small molecule complexes where the proteins have a maximum of 30% sequence similarity to those in our training data. Our results show that for NAD complexes (Fig. 2d) and other ligands (Fig. S1), the Semantic Vectors are significantly closer to the actual embeddings of protein B-NAD complexes than to randomly chosen protein-small molecule complexes (KS statistic 0.599, p-value < 0.001).

ATOMICA identifies residues on interaction interfaces involved in intermolecular bonds.

We also examine the ability of ATOMICA to differentiate between blocks which are involved in intermolecular bonds (hydrogen bonding, pi-stacking, hydrophobic interactions) and other blocks present at the interface (Fig. 2e). To probe the importance of a block to the embedding of the interaction complex, we define ATOMICASCORE for block i as the cosine similarity between the ATOMICA embedding of the graph of the interaction complex with block i masked out and the ATOMICA embedding of the graph without any masking (Fig. 2f). The blocks are ordered by ATOMICASCORE from lowest to highest, which reflects that blocks that are most important to the representation are at the start of the list. Although ATOMICA is not trained on any labels related to intermolecular bonds, this setup enables a zero-shot probing of its learned representations. We determine the blocks involved in intermolecular bonds using a protein structure profiler [80] for all protein-small molecule structures in the test set. Protein binding sites often contain conserved sequences [81, 82]. We therefore compare ATOMICA to a protein language model, which can predict the effects of mutations on protein fitness in a zero-shot manner [83], to assess whether such models can identify amino acids essential for small molecule binding. 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 retrieved, followed by ESM-2 with 2.4, and the random reference with 2.0. ATOMICA's ability to identify intermolecular bonds is also uniformly true across the most common types of intermolecular bonds observed (Fig. S2).

ATOMICA achieves generalizable interface representations across molecular modalities.

Although proteins, nucleic acids, ions, and small molecules are chemically distinct, their interactions are governed by similar intermolecular forces, including hydrogen bonds, aromatic

stacking, and van der Waals interactions. Generalizable representations of interaction interfaces across molecular modalities have not been demonstrated across ions, small molecules, proteins, and nucleic acids. 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. 2h). We reserve a test set of interface complexes with minimal sequence similarity and small molecular structure similarity to structures observed in training and validation to evaluate the models on (further details available at Online Methods 6). 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.

Pretraining across molecular modalities of interacting molecules improves block identity recovery over pretraining on a singular type of interacting molecules (Fig. 2i). ATOMICA trained across modalities improves masked block identity recovery on test set protein-DNA interactions from 0.24 to 0.71 AUPRC, a 190% improvement, with 2,253 training protein-DNA complexes. Similar gains in performance are observed for test set protein-RNA interactions from 0.19 to 0.55 AUPRC, where there are 2,975 training protein-RNA complexes. We also see an improvement from 0.32 to 0.67 AUPRC for test set protein-peptide complexes, which has 7,346 training protein-peptide complexes. The improvement of recovery on multiple modalities compared to single molecular modalities is observed uniformly across a full range of residue burial (Fig. S3), showing training with ATOMICA on multiple interacting modalities improves masked block identity prediction over training on a single interaction type.

ATOMICA's performance gains are correlated with dataset size (Fig. 2k), reflecting established scaling laws in LLMs where performance improves with dataset size [84]. Stratifying performance by intermolecular context, which we define by average distance to nearest intermolecular blocks, we observe training on all pairs of modalities improves performance across intermolecular context. Pretraining across all pairs of interacting modalities also yields a larger increase in performance for the blocks with less intermolecular context – for the blocks closest to other intermolecular blocks (8.0-8.4 Å away from 8 closest intermolecular blocks) there is a 13% improvement in AUPRC compared to a 28% improvement in AUPRC for blocks more distant to other intermolecular blocks (10.8-11.2 Å away from 8 closest intermolecular blocks) (Fig. S3b). As performance gains scale with dataset size increase, 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.

Building interfaceome networks for five molecular modalities.

Complex diseases are caused by a signaling network's dysregulation rather than a single protein [85] and often involve dysfunctional interactions with ions [86], small molecules [87, 88], lipids [89], or nucleic acids [90]. A disease pathway denotes the proteins involved in the disease. Proteins in a disease pathway tend to be clustered in the same network neighborhood [85], where network relations are defined via protein-protein interactions and maps of cellular pathways [91, 92]. 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 show proteins implicated in the same disease are more likely to share similar ion, small molecule, nucleic acid, and lipid interactions. Complementary to protein structure and sequence ML-based approaches for studying pathogenic variants in one protein sequence [93–95], ATOMICA considers the broader context of disease pathways in which the protein participates, capturing the complexity of disease mechanisms.

We explore the human interfaceome [35], defined as the set of human protein interfaces that mediate interactions with other molecules, including ions, small molecules, nucleic acids, lipids, and proteins. Since determining the structure of all molecular complexes involving proteins interacting with all other modalities in humans is still intractable, we focus on the protein interfaces to these modalities. To construct a network with protein interfaces as nodes, we first extract protein binding sites from each protein with ions, small molecules, nucleic acids, lipids, and proteins across 23,391 protein structures predicted by AlphaFold2 in the human proteome. We finetune ATOMICA-Interface from ATOMICA to support embedding of protein interfaces instead of complexes (refer to details in Online Methods 4). To ensure we are working with high-quality protein structures, we remove binding sites of protein interfaces that have low confidence pLDDT scores < 70. One ATOMICANET is constructed for each interacting modality, where nodes are proteins and edges connect pairs with high ATOMICA similarity between their modality-specific binding interfaces. The number of proteins embedded along with the size of the largest connected component (LCC) is as follows: 6,458 embeddings of proteins with ion interacting interfaces (yielding an *ion-based interfaceome network* with LCC of size 5,831 proteins; ATOMICANET-

Ion), 5,856 embeddings of proteins with small-molecule interacting interfaces (yielding *a small molecule-based interfaceome network* with LCC of size 5,246; ATOMICANET-Small-Molecule), 6,649 embeddings of proteins with nucleic acid interacting interfaces (yielding *a nucleic acid-based interfaceome network* with LCC of size 5,974; ATOMICANET-Nucleic-Acid), 6,766 embeddings of proteins with lipid interacting interfaces (yielding *a lipid-based interfaceome network* of size 6,055; ATOMICANET-Lipid), and 17,158 embeddings of proteins with protein interacting interfaces (a *protein-based interfaceome network* of size 15,450; ATOMICANET-Protein) (Fig 3a). Further details for network construction are available in Online Methods 4.

Next we explore the pathways observed in the network across 82 diseases and their disease-associated proteins. For each disease, we determine the disease-associated proteins with Open-Targets [96], and overlay these proteins onto the five ATOMICANETs. 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 3b). The edges in the ATOMICANETs are also distinct from any existing protein network, as they capture similarity in protein interaction interfaces. The average overlap in edges across the five ATOMICANETs with PPI networks is 2.7%, with co-occurence networks is 1.0%, neighbourhood networks is 0.3%, and across all sources of evidence is 6.4% [97]. To analyze the observability of disease pathways in ATOMICANETs we calculate the size of the largest disease pathway component, which is the proportion of disease-associated proteins in the largest connected component of the disease pathway [98]. The diseases selected are across disease areas ranging from cancer, immune system diseases, to nervous system diseases. Given that disease pathways are only detectable if there are at least 25 proteins in the network [85], we limit our analysis to diseases which have at least 25 proteins in a ATOMICANET. 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 3c). 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 interfaceome-based ATOMICANET networks.

To assess the significance of disease pathways in each ATOMICANET, we evaluated the size of their largest pathway components and compared them against a reference distribution (for details refer to Online Methods 4). This reference was generated by randomizing disease proteins

while preserving both pathway size and node degree distribution. We applied the Benjamini-Hochberg procedure to control the false discovery rate (FDR) and considered a p-value < 0.05 as statistically significant.

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 3d), 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 [99]. Proteins in these two components form key interactions with PIP2, a minority lipid component of the cell membrane [100].

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 3e), 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 [99]. 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 [101]. 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 3f), 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 [99]. 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 [102, 103], cardiac actin (ACTC1) – a crucial sarcomeric protein which is also strongly associated with the disease [104], and HRAS – a GTPase regulating a host of signaling pathways and cellular responses [105].

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 [85].

ATOMICANET associates lipid and ion interacting targets for autoimmune channelopathies.

We next test the ability of ATOMICANETs in associating disease proteins for autoimmune neurological diseases. Autoimmune peripheral neuropathy and multiple sclerosis are both autoimmune diseases that involve the immune system mistakenly attacking components of the nervous system, but they affect different parts: autoimmune peripheral neuropathy targets the peripheral nervous system, while multiple sclerosis primarily affects the central nervous system, specifically the myelin sheaths of nerve fibers. Due to the importance of ion signaling in neurons [106–108] and the association of lipid dysregulation with disease onset [109], ATOMICANET presents an exciting opportunity to analyze the association of disease proteins in a modality-specific manner. We use diffusion-based methods, which are a state-of-the-art approach for the detection of disease-associated proteins [110]. Random walkers from seed proteins are used to detect other disease-associated proteins based on which proteins were visited most frequently in the network. We perform 5-fold cross-validation, using four folds as seed disease proteins and holding out the fifth. This is repeated 100 times with different splits. A protein's hit rate is defined as the proportion of splits in which it appears among the most likely nodes visited by the random walk (see Online Methods 4).

For multiple sclerosis we identify 69 out of 119 disease proteins, and for autoimmune peripheral neuropathy we identify 40 out of 44 disease proteins that are present across the five networks. KCNA1, a voltage-gated potassium channel, was frequently associated in ATOMICANET-Ion with a hit rate of 100% and 100% for the respective diseases. KCNA2 was also nominated with

a hit rate of 100% and 100% for the respective diseases by ATOMICANET-Ion. These two Kv1 channels belong to a major subfamily of voltage-gated potassium channels that are expressed in axons throughout both the central and peripheral nervous systems, and the exposure of these channels during demyelination in multiple sclerosis impairs nerve conduction [111]. Given that these proteins are found in the cell membrane, we also expect to find ATOMICANET-Lipid to detect these proteins. This proves to be the case with KCNA1 (70% hit rate) and KCNA2 (100% hit rate) both frequently associated in ATOMICANET-Lipid for multiple sclerosis. Likewise, for autoimmune peripheral neuropathy, we see that KCNA1 has an 83% hit rate and KCNA2 has a 100% hit rate. We also find KCNQ3 to be highly associated (hit rate 100%) in ATOMICANET-Lipid for multiple sclerosis and autoimmune neuropathy. This target is known to be involved in the M current, a neuronal K+ conductance [107, 108]. Several sodium channel alpha subunit family proteins are also found to be associated with multiple sclerosis in ATOMICANET-Lipid. This includes SCN2A, SCN8A both with hit rates of 100%. These targets are shown to be over-expressed in multiple sclerosis [112]. With the importance of transmembrane ion channels in multiple sclerosis and autoimmune peripheral neuropathy, we show the ability of ATOMICANET to identify targets across interacting molecular modalities.

Identifying lymphoma-associated proteins across five molecular interaction modalities.

To demonstrate how complementary associations emerge from different ATOMICANET modalities, we analyze disease-associated proteins in lymphoma, a malignancy originating from lymphocytes, often involving lymph nodes and lymphatic tissues throughout the body. Comprehensive understanding of lymphoma requires dissecting interactions among proteins and interaction partners, including nucleic acids, lipids, ions, small molecules, and other proteins. Across the five modality-specific ATOMICANET networks, some disease-associated proteins are shared, with a mean pairwise Jaccard similarity of protein nodes in the networks of 0.24. Nonetheless, random walks on each network highlight a distinct subset of lymphoma-associated proteins from the total of 296 proteins analyzed (Fig. 3g).

In ATOMICANET-Nucleic-Acid, the analysis identifies transcriptional regulators and epigenetic modifiers. Frequently nominated proteins include CREBBP (hit rate 92%) and EP300 (90%), both histone acetyltransferases crucial for transcriptional regulation, as well as ETV6 (69%) and BCL6 (50%), transcription factors directly implicated in lymphoid cell differentiation and lymphoma pathogenesis [113]. In ATOMICANET-Lipid, identified proteins emphasize signaling receptors and nuclear hormone receptors. Prominent examples are the FLT3 cell surface tyrosine

protein kinase receptors (hit rate 93%) and KIT (81%), both involved in critical growth factor signaling cascades in hematologic malignancies [114]. Additionally, the nuclear retinoic acid receptors RXR-$\alpha/\beta/\gamma$ emerge prominently, with hit rates of 62%, 62%, and 58%, respectively, reflecting roles in gene regulation through lipid-mediated signaling. In ATOMICANET-Ion, the analysis highlights zinc finger domain-containing proteins involved in transcriptional and DNA replication processes. These include retinoic acid receptors RXR-$\alpha/\beta/\gamma$ (each with a hit rate of 94%), again emphasizing their dual roles in ion- and lipid-binding contexts. Other prominent targets are catalytic subunits of DNA polymerase complexes, POLD1 (hit rate 87%) and POLE (73%), essential in DNA replication fidelity and genome stability. Other proteins identified include BCL11B (55%), a critical transcription factor regulating T-lymphocyte differentiation and survival, and WT1 (48%), known for its roles in cellular proliferation and apoptosis. In ATOMICANET-Small-Molecule, nominated proteins are primarily metabolic and chromatin remodeling enzymes. Key examples include IDH2 (hit rate 95%), an enzyme involved in NADP-dependent metabolic pathways frequently mutated in hematological malignancies [115]. We also observe ATP-dependent chromatin remodelers such as SMARCA (hit rate 92%), known to activate BRG1-dependent transcriptional complexes, and CHD2 (80%), a DNA-binding helicase implicated in chromatin remodeling and genome integrity maintenance [116]. Finally, in ATOMICANET-Protein, proteins integral to protein complexes involved in cellular signaling and gene regulation are identified. These include STAT3 (hit rate 78%), a transcription activator functioning via homo- or hetero-dimerization with STAT1 [117], and TBL1XR1 (76%), a component of nuclear receptor corepressor complexes that recruits ubiquitin-proteasome machinery for targeted protein degradation [118]. These results illustrate how each molecular modality captures distinct yet complementary aspects of lymphoma biology, spanning cell surface receptor signaling, chromatin regulation, metabolic enzymatic pathways, DNA replication and repair, and protein complex-mediated transcriptional regulation.

Revealing functional roles in the dark proteome through systematic annotation of ion and cofactor binding sites.

Ion and cofactor binding sites are conserved functional features widely distributed throughout the proteome [82, 119]. Even among proteins that differ significantly in their overall sequence or structure, ligand-binding regions often align with functional domains preserved across evolution [119]. 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 [120], suggesting the potential of structure-based approaches, such as ATOMICA, to anno-

tate functional binding sites in regions of the proteome that currently lack any functional description, collectively known as the *dark proteome* [59–61].

Advances in protein structure prediction (AlphaFold-derived databases [121], such as AFDB) enable systematic analysis of structural and evolutionary relationships across nearly the entire known protein universe [60, 122]. These analyses have identified substantial regions of the protein sequence space that are distinct in both structure and evolutionary history yet lack functional characterization. Such structurally distinct but functionally unannotated groups of proteins are known collectively as *dark clusters* [60]. Currently, 711,705 protein clusters (30.9% of all clusters cataloged in the AlphaFold (AFDB) FoldSeek) fall into this dark category (Fig. 4a). Proteins in these dark clusters frequently exhibit novel structural folds that span all major domains of life, including Bacteria, Archaea, and Eukarya. Some proteins are ancient, representing deep branches on the evolutionary tree. Investigating these dark clusters provides opportunities to discover previously unknown protein functions, elucidate novel molecular mechanisms, and reconstruct evolutionary trajectories leading to present-day protein diversity [59, 123, 124].

Because 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 [125]. ATOMICA addresses this challenge by structurally annotating ion and cofactor binding sites in the dark proteome. By modeling the geometric and chemical characteristics of binding sites, ATOMICA generalizes beyond proteins closely related to those encountered during training (Fig. 4a).

We use ATOMICA to systematically functionally annotate proteins in dark clusters in the proteome. It has been observed that some of these unannotated proteins have predicted structures of poor quality, so we restrict our analysis to protein clusters with high-confidence AF2 structures (pLDDT > 90) which are 33,482 clusters and we choose the representative AF2 member of each cluster for further investigation. Ion and ligand binding sites are identified on the surface of the protein with a deep learning-based classifier, PeSTo. In total, 2,851 proteins are identified with ion binding sites and 969 proteins are identified with ligand binding sites. We finetune ATOMICA to predict ion and ligand identities given respective protein pockets from structures in the PDB for 9 metal ions and 12 commonly found cofactors (Online Methods 5). We denote these models ATOMICA-Ligand (Fig. 4b) and apply the finetuned models to the predicted binding sites of the dark proteome and annotate 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 visualize the bound structure of

the proteins with their predicted ions and ligands. We confirm the quality of our predictions with AlphaFold3 ipTM scores of the complexes, which serve as a quantitative metric for the generation quality of complexes [126, 127]. 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. 4c). 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.

Discovering conserved ion and cofactor binding sites reveals functions in the dark proteome. ATOMICA-Ligand assigns a high score (0.76) for Mg^{2+} to an ion binding site in A0A0M0-BG38 (Fig. 4d). This protein represents a cluster of 45 members that have the lowest common ancestor of cellular organisms. Sequence-based prediction method ProtNLM [128], suggests this is a phosphatidate cytidylyltransferase (score 0.82), an enzyme that combines cytidine triphosphate (CTP) and phosphatidate to produce CDP-diacylglycerol (CDP-DAG). These proteins are Mg^{2+}-dependent enzymes for coordination with phosphorous groups and in agreement with the ATOMICA-Ligand annotation of Mg^{2+} binding site.

For two uncharacterized Bacteria protein clusters, ATOMICA-Ligand annotates high confidence Zn^{2+} binding motifs which share similarity to metallopeptidases. A0A6C0E0P4 (Fig. 4e) is a representative protein of a cluster of size 25 and originates from viral metagenome sequencing, other cluster members share a lowest common ancestor of alphaproteobacteria. A0A7Y4VG56 (Fig. 4f) is a representative protein of a cluster of size 18 and originates from Ignavibacteria bacterium, the lowest common ancestor of the cluster is Bacteria. The Zn^{2+} -binding site in both proteins consists of two histidines and one glutamate, where the histidines are at positions i and $i + 4$ on the same α-helix, while the glutamate is contributed by a separate α-helix, and for both proteins ATOMICA-Ligand predicts Zn^{2+} with a score of 0.93. This Zn^{2+} binding HEXXH motif, which is present in most metallopeptidases, specifically in zincins it is comprised of two histidines located on the active-site helix and positioned between a beta sheet with at least three strands and a backing helix with the glutamate residue [129, 130]. Structural alignment of the proteins with thermosylin (PDB: 2TLX) reveals alignment of the Zn^{2+} binding motifs, and a TM-align score of 0.49 and 0.44 for A0A6C0E0P4 and A0A7Y4VG56, respectively. This is also in agreement with sequence-based functional annotation of these proteins which suggests peptidase domain containing protein, albeit with relatively low confidence of 0.12 and 0.21 for A0A6C0E0P4

and A0A7Y4VG56, respectively.

ATOMICA-Ligand further identifies for an ion binding site, belonging to a protein in a cluster of previously uncharacterized Bacteria proteins, to be a C4 zinc finger motif. While zinc fingers are frequently discovered in eukaryotes, they are less characterized in Bacteria [131]. W8VVQ2 is given a score of 0.98 by ATOMICA-Ligand for a Zn^{2+} binding site. W8VVQ2 is part of a cluster with 66 members from Nonlabens marinus isolated from the Pacific Ocean [132] and this cluster has a lowest common ancestor of cellular organisms. Thus, it is exciting that we identify in this cluster, with ancient origins, a total of 61 Bacteria C4 zinc fingers out of 66 in the cluster. The other 5 proteins in the cluster also contain zinc fingers, albeit with a more unconventional 3Cys-1Glu motif. Coupling this with ProtNLM predictions, it suggests HNH endonuclease with a score of 0.47, which features a $\beta\beta\alpha$-metal topology [133]. We identify this structure in Fig. 4g and the conserved histidine and aspartic acid of this motif [134]. While HNH endonucleases bear topological similarity to zinc fingers, they do not have zinc fingers. Hence, through a focus on annotating ligands to dark proteins, ATOMICA has identified a putative novel class of bacterial C4 zinc fingers with endonuclease function.

Finally, ATOMICA-Ligand identifies heme for binding sites in two large ancient clusters which are part of a functional site of putative transmembrane cytochrome subunits. A0A6B2YTG2 is predicted to bind to heme with a score of 0.76 by ATOMICA-Ligand (Fig. 4h). A0A6B2YTG2 is part of a cluster of size 134, from an unclassified strain of Streptomyces. The lowest common ancestor of this cluster is cellular organisms, and some members of the cluster are annotated as membrane proteins. It is closely related to cluster A0A2V8JCM5 (TM-align score 0.911), comprised mostly of uncharacterized proteins and has 113 cluster members. ProtNLM assigns domains of unknown function to A0A6B2YTG2, but for A0A0K3BHV0 a low confidence prediction of a copper resistance D domain was made. However, PeSTo does not find any high confidence metal ion binding sites for A0A0K3BHV0. Instead, ATOMICA-Ligand predicts the A0A0K3BHV0 small molecule binding site to bind to heme with a score of 0.33. We find further evidence to support heme-binding with A0A7X0NTZ6, a member of the A0A0K3BHV0 cluster, a succinate dehydrogenase/fumarate reductase cytochrome b subunit protein where heme binding is necessary for electron transfer function. We see similarities in the heme binding motif in A0A6B2YTG2 and that of cytochrome b, where heme is bound non-covalently via histidine [135]. A more distant cluster containing protein A0A2T0LHX9 and five other members further supports the heme-binding cytochrome annotation, as this cluster is annotated as cytochrome c/quinol oxidase subunit I. This

protein shares high sequence similarity with A0A6B2YTG2 (TM-align score 0.715). Therefore, through annotation of heme binding sites with ATOMICA-Ligand, the model characterizes a cluster of ancient origin to be putative cytochrome proteins.

Discussion

ATOMICA is a deep learning 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. These embeddings recover functional properties such as identifying residues involved in binding, and they stratify proteins by interaction profiles that are predictive of disease relevance. We construct ATOMICANETs, which are modality-specific interfaceome networks that capture disease modules in 27 conditions and allow the identification of disease-associated proteins in autoimmune neuropathies and lymphoma. By leveraging ATOMICANETs to focus on diverse interface modalities, we gain an orthogonal, more holistic view of disease mechanisms than any single ATOMICANET alone could provide. ATOMICA generalizes to previously uncharacterized proteins in the dark proteome, allowing annotation of ion and cofactor binding sites in structurally and functionally novel protein families.

A key limitation of ATOMICA is its reliance on high-resolution molecular structures. Experimental structures are not uniformly available across all molecular complexes due to the challenges of crystallization, conformational flexibility, and molecular disorder. However, recent advances in structure prediction, such as AlphaFold3 [16] and RoseTTAFold [17], can generate high-confidence structures across a wide range of biomolecular modalities, including protein-ligand and protein-nucleic acid interactions. These tools can help extend the applicability of ATOMICA to protein regions that lack experimental coverage. Nonetheless, some structural classes remain underrepresented. Non-globular domains, particularly intrinsically disordered regions (IDRs), are not well captured in current structural databases due to their inherent flexibility [136]. Yet, many IDRs undergo disorder-to-order transitions upon binding and play essential roles in cellular regulation and signaling [137]. Similarly, flexible regions such as antibody complementarity-determining regions are critical for antigen recognition but are difficult to resolve accurately in structure prediction [138]. In these cases, relying solely on structural input limits the accuracy and coverage of interaction modeling. ATOMICA could be extended to integrate sequence-based features, which have proven informative for IDRs [139], and may improve representation quality in structurally

ambiguous regions.

Another limitation is the relatively small size of the structural dataset compared to high-throughput experimental interaction screens. Methods such as yeast two-hybrid [140], affinity purification mass spectrometry [141], DNA-encoded libraries [142], proximity ligation assays [143], fluorescence resonance energy transfer [144], and surface plasmon resonance [145] generate large-scale data on molecular interactions without requiring detailed 3D structures. These techniques often capture transient or weak interactions that are difficult to observe structurally. Incorporating such non-structural data would broaden the applicability of ATOMICA. However, doing so would require generating plausible structural models for the molecular pairs or developing multi-view architectures that can jointly learn from sequence, structure, and interaction data [146, 147].

ATOMICA provides a general framework for reasoning about biomolecular interfaces, which opens several avenues for future applications. Its ability to represent interaction types with shared geometric and chemical principles suggests use in unifying datasets across structural biology, chemical biology, and systems biology. For example, ATOMICA embeddings can be integrated into models that predict the effects of sequence variation on binding interfaces, helping to interpret variants in clinical and population-scale sequencing studies. The compositional structure of the latent space enables interface-level comparison and retrieval, which could support the repurposing of ligands, cofactors, or binding domains across targets. ATOMICA can serve as a foundation for generative models that design interacting molecules, such as small molecule inhibitors, peptides, or nucleic acid aptamers, based on the geometry of the interface rather than the global structure of the protein. Finally, by embedding protein interfaces independent of specific ligands, ATOMICA could facilitate multitarget optimization or selectivity profiling in therapeutic design.

Understanding molecular interactions is essential to uncover the principles of cellular organization and disease mechanisms. Foundation models represent molecules in isolation or restrict the modeling to a narrow range of interaction types. ATOMICA addresses this limitation by learning unified representations of interaction interfaces across molecular modalities. As efforts toward a virtual cell intensify [148], 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.

Data availability. Datasets are available on Harvard Dataverse at https://doi.org/10.7910/DVN/4DUBJX.

Code availability. The code to reproduce the results, documentation and usage examples is available on GitHub at https://github.com/mims-harvard/ATOMICA. The model weights are available at https://huggingface.co/ada-f/ATOMICA. Details are also available at https://zitniklab.hms.harvard.edu/projects/ATOMICA.

Acknowledgements. We thank Michael Desgagne and Michael Sun for their helpful discussions on the manuscript. A.F. is supported by the Kempner Graduate Fellowship at Harvard University. We gratefully acknowledge the support of NIH R01-HD108794, NSF CAREER 2339524, US DoD FA8702-15-D-0001, awards from Harvard Data Science Initiative, Amazon Faculty Research, Google Research Scholar Program, AstraZeneca Research, Roche Alliance with Distinguished Scientists, Sanofi iDEA-iTECH, Pfizer Research, Chan Zuckerberg Initiative, John and Virginia Kaneb Fellowship at Harvard Medical School, Biswas Computational Biology Initiative in partnership with the Milken Institute, Harvard Medical School Dean's Innovation Fund for the Use of Artificial Intelligence, and Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

Authors contributions. A.F. developed and implemented the ATOMICA model, conducted extensive benchmarking, and performed detailed analyses of its performance and applications. A.F. and M.Z. designed the study. Z.Z. contributed to the interpretability framework and developed the ATOMICAScore analysis. A.Z. assisted with the analysis of disease-associated proteins and the evaluation of ATOMICANet. All authors contributed to writing and editing the manuscript.

Competing interests. The authors declare no competing interests.

Figures

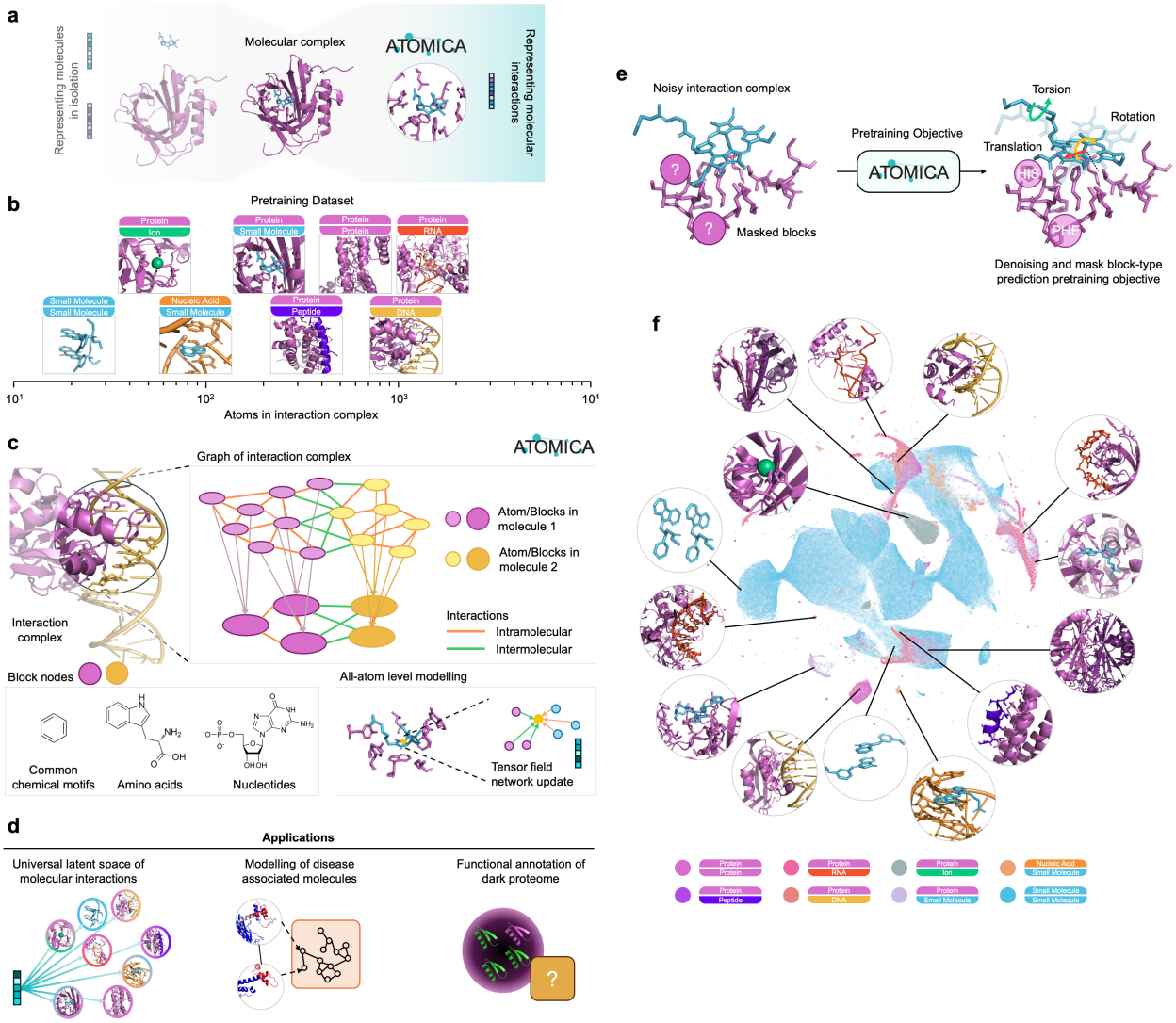

Figure 1: Overview of ATOMICA pretraining data, architecture, and latent space. **a** ATOMICA focuses on representing molecular interactions which is complementary to models representing individual molecules. **b** Input modalities and the number of atoms in the interaction complexes that ATOMICA are pretrained on. **c** 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. **d** ATOMICA enables modeling of a universal latent space of intermolecular interactions. This is leveraged for modeling of disease associated molecules and functional annotation of the dark proteome. **e** Self-supervised pretraining objectives for generating ATOMICA embeddings. **f** UMAP of latent space of all interaction complexes seen during pretraining.

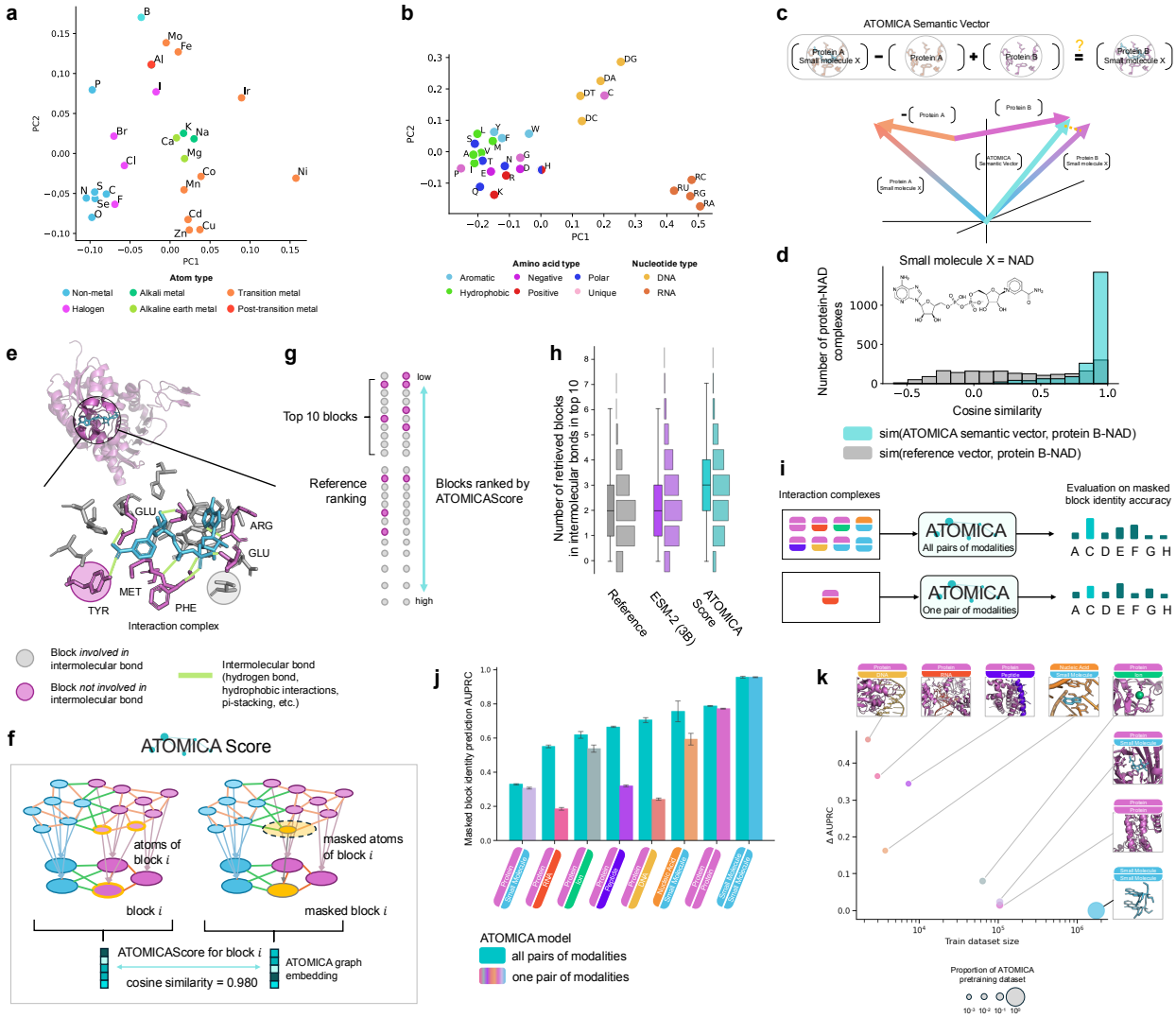

Figure 2: Analysing latent space of ATOMICA. **a** Principle components of mean embedding of elements in the pretraining validation and test set. **b** Principle components of mean embedding of amino acids and nucleotides in the pretraining validation and test set. **c** Compositional algebra of embeddings in the latent space of ATOMICA. **d** Similarity of ATOMICA Semantic Vectors to protein-NAD embeddings. Reference vectors are the embedding of a protein-small molecule complex chosen at random. **e** Some blocks in the interaction complex are involved in intermolecular bonds (hydrogen bonds, hydrophobic interactions, pi-stacking, etc.). **f** Definition of ATOMICAScore for block i, cosine similarity between interaction graph with block i masked and unmasked. **g** 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. **h** 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. ATOMICA nominates more amino acids involved in intermolecular bonds than ESM-2 and the reference. **i** Schema to test transferability 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. **j** Masked block identity AUPRC for different pairs of modalities for ATOMICA trained on all pairs of modalities and ATOMICA trained on one pair of modalities. ATOMICA trained on all pairs of modalities outperforms ATOMICA trained on one pair of modalities. **k** Increase in AUPRC between training on one pair of modalities and all pairs of modalities compared to dataset size of the one pair of modalities. Performance gains scale with dataset size increase, with largest improvements in AUPRC observed for pairs of modalities with the least data.

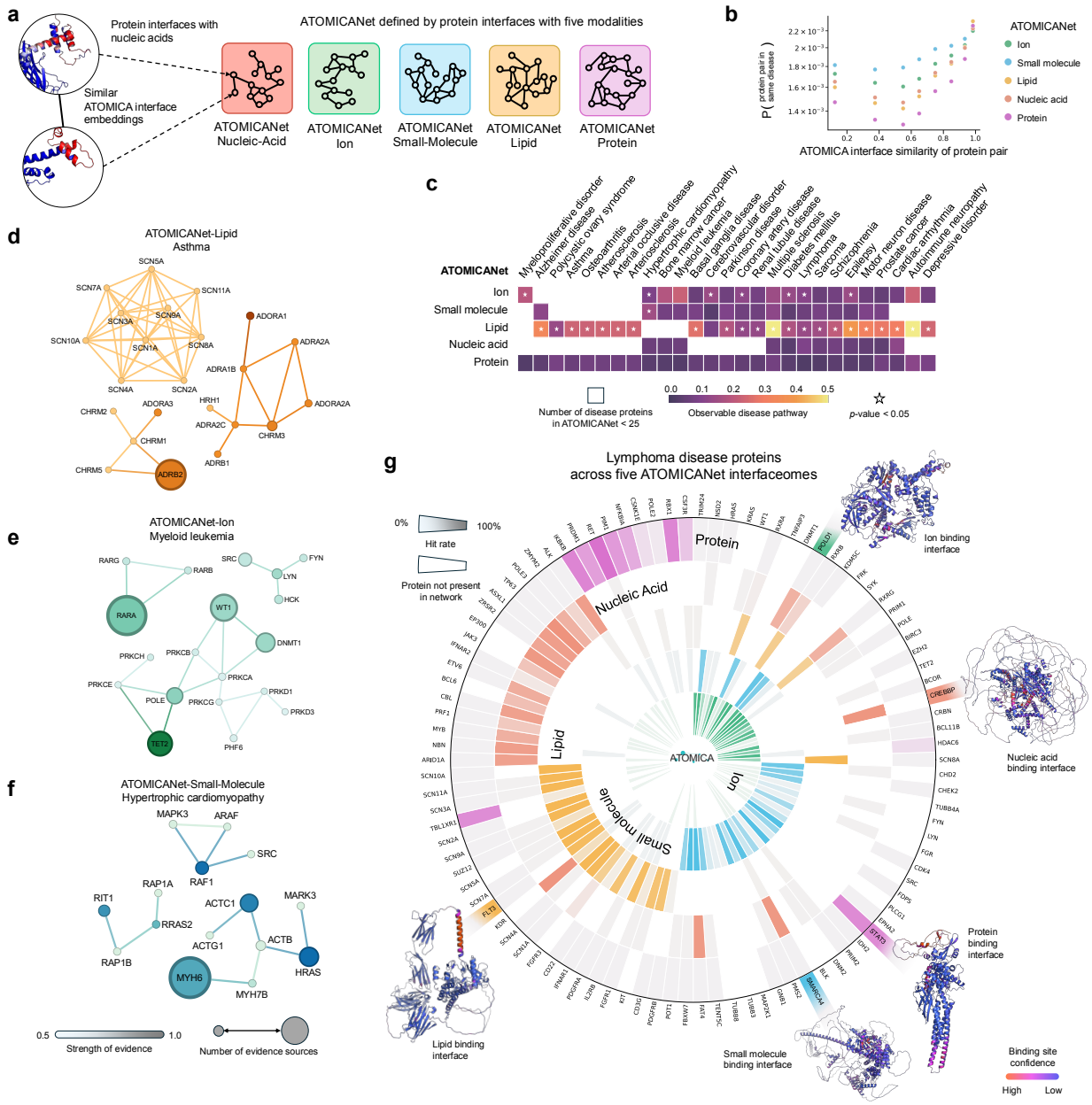

Figure 3: Interfaceome disease pathways on ATOMICANETS. **a** Set up of the modality specific networks based on ATOMICA embedding similarity of protein interfaces with ions, small molecules, lipids, nucleic acids, and proteins. **b** Cosine similarity of ATOMICA embeddings of protein interface pairs across five interacting modalities compared to the probability of the protein pair being involved in the same disease. **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. The three largest pathway components for: **d** asthma in ATOMICANET-Lipid, **e** myeloid leukemia in ATOMICANET-Ion, **f** hypertrophic cardiomyopathy in ATOMICANET-Small-Molecule. **g** Detected proteins using random walk across all five ATOMICANETS for lymphoma. Only the disease proteins which appear in at least one ATOMICANET are visualized.

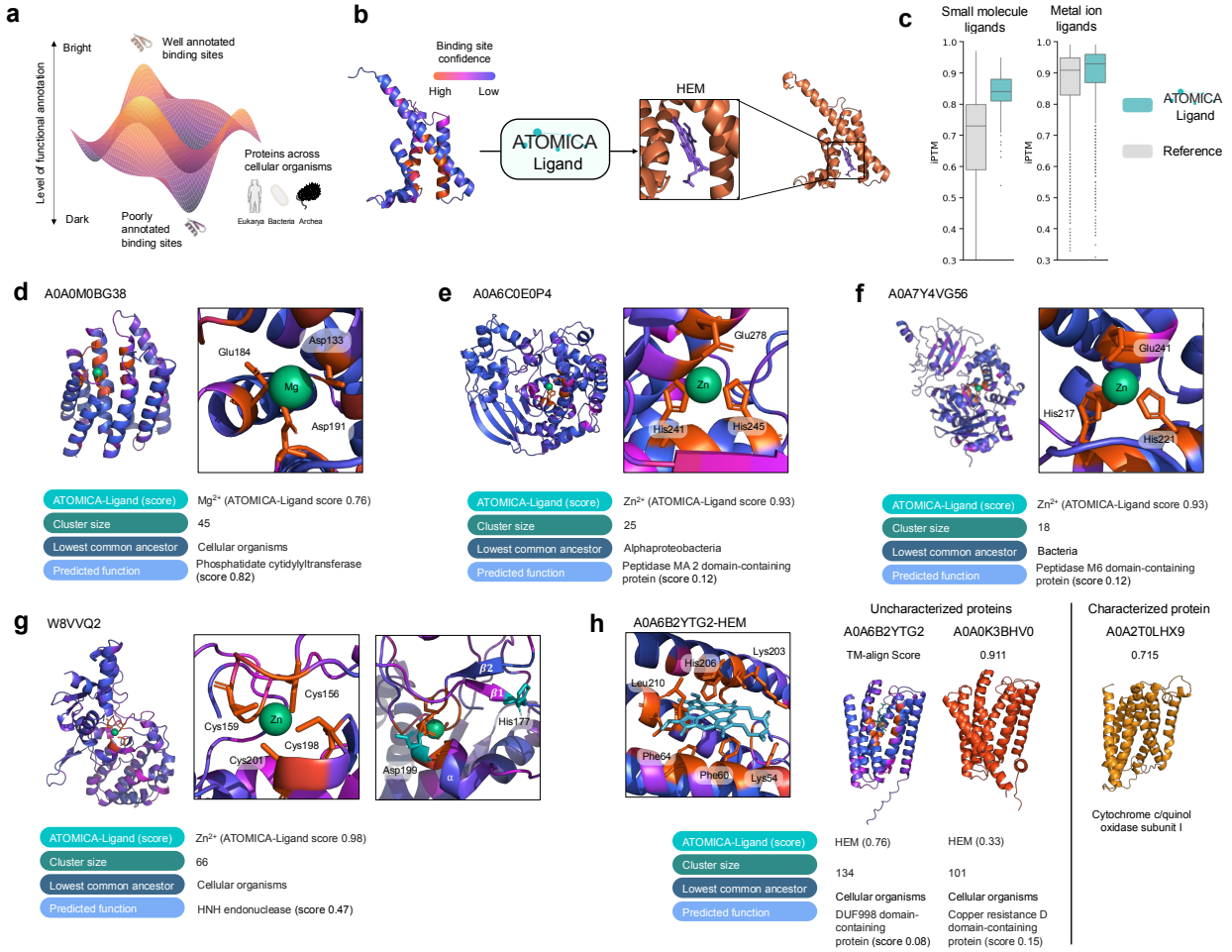

Figure 4: Dark proteome annotation with ATOMICA. a Annotation of binding sites for proteins across cellular organisms. **b** Prediction of ligands for metal ion and small molecule binding sites of proteins in the dark proteome. **c** AlphaFold3 ipTM scores of complexes from ATOMICA-Ligand annotated small molecule and metal ion compared to reference. Visualization of AlphaFold3-predicted structures for protein-ligand annotations made by ATOMICA-Ligand. Cluster size and lowest common ancestor are derived from Foldseek clustering of the AlphaFold Protein Structure Database [60]. Predicted functions are obtained from ProtNLM, a sequence-based protein function prediction model [128]. The following ATOMICA-Ligand annotations are shown: **d** Mg^{2+} annotation for A0A0M0BG38, a putative phosphatidate cytidylyltransferase, where Mg^{2+} binding is required for enzymatic activity. **e** A0A6C0E0P4 and **f** A0A7Y4VG56 are annotated with Zn^{2+} and predicted to be metallopeptidases. Both contain the conserved HEXXH Zn^{2+}-binding motif characteristic of this family. **g** Zn^{2+} annotation for W8VVQ2, with the predicted binding site indicating a C4 zinc finger domain, a motif that remains undercharacterized in Bacteria. **h** Heme annotation for A0A6B2YTG2, with comparison to neighboring clusters A0A0K3BHV0 and A0A2T0LHX9 suggesting these uncharacterized clusters may contain putative cytochrome proteins.

Online Methods

The Methods describe (1) the curation of datasets, (2) hierarchical all-atom graph neural network architecture of ATOMICA, (3) training ATOMICA, (4) studying disease pathways with ATOMICA, (5) the fine-tuning of ATOMICA for ion and ligand pockets, (6) metrics and statistical analyses.

1 Curation of datasets

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 [149], we use a one-hot encoding of the molecular complex from a vocabulary of 290 common chemical motifs [64] 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.

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 protein-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 [150]. 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.

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 [64]. 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.

2 Hierarchical all-atom graph neural network

Overview. 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.

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}.

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 [66, 151–153]. We build on the E(3)-equivariant neural network layers presented by Tensor-Field Networks implemented in e3nn [154] and DiffDock [10]. 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 $\mathbf{h}^{\text{atom}}_{a,(\lambda,p)}$ 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:

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

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

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.

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.

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 $\mathbf{h}_i^{\text{graph}}$.

Self-supervised learning with denoising and block identity masking. 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 [155], and it has proven that it is related to learning a force field of per-atom forces [25, 156]. In addition, denoising is linked to score-matching which has also been popular in training generative models [10, 157] as well as unsupervised binding affinity prediction [73]. Thus, this motivates the application of denoising as an objective for self-supervised training.

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 at uniform random:

- 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.

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 [10]:

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

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 $\mathcal{N}_z = \{a \mid a \in \mathcal{A}', ||\widehat{r}_{za}|| < 5\mathring{A}\}$ which is all atoms in the perturbed molecule within 5 Å to the center of the bond.

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

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, $Y^2(\widehat{\mathbf{r}}_z)$, and the unit direction vector between the center of the bond and atom a, $Y(\widehat{\mathbf{r}}_{za})$. 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 [10, 73].

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:

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

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

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

.

3 Training details for ATOMICA

Overview. 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.

Hyperparameter tuning. We employed a hyperparameter optimization strategy utilizing Ray Tune [158] in conjunction with Optuna [159] and the Asynchronous Successive Halving Algorithm (ASHA) scheduler [160]. 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.

Implementation. ATOMICA is implemented with PyTorch (Version 2.1.1) [161] and PyTorch Geometric (Version 2.1.1) [162]. Training runs were monitored with Weights and Biases [163]. Models are trained on 4 NVIDIA H100 Tensor Core GPUs in parallel.

4 Analyses of disease pathways with ATOMICA

Overview. Following pretraining ATOMICA, we finetune the learned representations of the molecular complexes for diverse biomedical downstream tasks. We demonstrate ATOMICA's

ability to fingerprint protein surfaces to a diverse set of binders: ions, ligands, proteins, nucleic acids, and lipids. We demonstrate the versatility of these embeddings by studying the presence of disease pathways in networks constructed from these embeddings, showing similar protein interfaces across interacting modalities are likely to be involved in the same disease.

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:

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

where sim is cosine similarity and τ is the temperature factor. This contrastive loss is adapted from the normalized temperature-scaled cross-entropy loss [164]. 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.

Detection of binding sites across the human proteome with PeSTo. We employ PeSTo (Version 4.1) [35], 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 [6, 121]. 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.

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 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 [165]. 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. Visualisations of the networks are constructed with Gephi [166].

Therapeutic targets dataset. We extract targets for diseases from Open Targets (2024-09) [99]. 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. The list of 82 diseases can be found in Supplementary Table 1.

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 [85], 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 [98], 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.

Disease pathway detection in ATOMICANETs. Given ATOMICANETs and the disease proteins, the task is to predict new proteins that are likely associated with the disease. This has

been widely explored on PPI networks, with diffusion-based methods demonstrating strong performance [98, 167]. We use a variant of random walks with restarts: a random walker starts from a given seed protein, and at every time step, it can move to neighboring nodes or restart from a seed node. The probability of a walker being at node i at time t is then given by $\mathbf{p}^t = (1-r)\mathbf{A}\mathbf{p}^{t-1} + r\mathbf{p}^0$, where $\mathbf{p}^0$ is the initial probability, which is equal probability to start from each seed node, $r = 0.75$ is the restart probability, $\mathbf{A}$ is the normalized adjacency matrix. We run the random walk until $||\mathbf{p}^t - \mathbf{p}^{t-1}|| < 10^{-6}$ and $t < 1000$. The nominated proteins are then the 500 proteins with the highest probability of being visited. For each disease, modality, and random seed, we set up a 5-fold cross-validation split on V_d^{modality}; for each fold, we set them to be disease proteins to be detected, and the proteins in the remaining folds are set as seed nodes. This is repeated for 100 independent seeds, and the hit rate is calculated as the proportion of seeds for which the disease protein was detected.

5 Dark proteome binding site characterization with ATOMICA-Ligand

Overview. 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.

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 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.

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 [150]. 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).

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 [60]. 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 [128].

6 Metrics and statistical analyses ATOMICA

Visualization of embeddings. We embed 2,105,459 molecular interaction complexes from training, validation, and testing datasets, and project these embeddings into two dimensions using Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [168] with 30

neighbors and a minimum distance of 0.001. All interaction complexes were visualized with PyMOL [169].

Visualization of atom and block identity. From molecular interaction complexes in training, validation, and testing datasets, we take the element-wise average atom and block embeddings for each identity. For the atom and block embeddings, we projected the embeddings into two dimensions using Principal Component Analysis (PCA).

Composition of ATOMICA embeddings. We analyze the composition of protein-small molecule embeddings for interaction complexes in the test set. To get the embedding of the pocket only, we remove the atom and block nodes corresponding to the small molecule and embed the graph with ATOMICA. We evaluated our algebraic composition method by measuring the cosine similarity between ATOMICA Semantic Vectors and their corresponding original complexes. For reference, we compared this similarity against a reference distribution obtained by calculating the cosine similarity between the original complexes and embeddings of randomly sampled protein-small molecule complexes from the test set. This comparison framework allows us to assess whether our composition method produces embeddings that meaningfully capture the properties of the actual complexes. This analysis was run for the following test set small molecule complexes: NAD (218 complexes), FAD (215 complexes), HEM (168 complexes), PLP (157 complexes), and ATP (105 complexes).

Capturing intermolecular bonds with ATOMICASCORE. To identify amino acids involved in intermolecular interactions, we analyzed protein-small molecule complexes in the test set using PLIP [80]. We then quantified the contribution of each amino acid to the interaction by calculating an importance score, ATOMICASCORE. 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. Within the 10 amino acids with the lowest ATOMICASCOREs we count the number of amino acids involved in intermolecular bonds. For reference, we compare this to randomly nominating 10 amino acids at the interface, and to ESM-2 (3 billion parameters) [19] and score amino acids by the log likelihood of the masked mutant compared to the original amino acid [83].

Training single modality ATOMICA. 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.

References

1. Stärk, H., Ganea, O.-E., Pattanaik, L., Barzilay, R. & Jaakkola, T. EquiBind: Geometric deep learning for drug binding structure prediction. *arXiv preprint arXiv:2202.05146* (2022).

2. Ganea, O.-E. *et al.* Independent se (3)-equivariant models for end-to-end rigid protein docking. *arXiv preprint arXiv:2111.07786* (2021).

3. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

4. Zhang, Z., Shen, W., Liu, Q. & Zitnik, M. Pocketgen: Generating full-atom ligand-binding protein pockets. *arXiv* (2024).

5. Zhang, Y., Huang, W., Wei, Z., Yuan, Y. & Ding, Z. Equipocket: an e (3)-equivariant geometric graph neural network for ligand binding site prediction. *arXiv preprint arXiv:2302.12177* (2023).

6. Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).

7. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

8. Wu, R. *et al.* High-resolution de novo structure prediction from primary sequence. *BioRxiv* 2022–07 (2022).

9. Mirdita, M. *et al.* Colabfold: making protein folding accessible to all. *Nature methods* **19**, 679–682 (2022).

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

11. Pei, Q. *et al.* Fabind: Fast and accurate protein-ligand binding. *Advances in Neural Information Processing Systems* **36** (2024).

12. Zhang, X. *et al.* Efficient and accurate large library ligand docking with karmadock. *Nature Computational Science* **3**, 789–804 (2023).

13. Cao, L. *et al.* Design of protein-binding proteins from the target structure alone. *Nature* **605**, 551–560 (2022).

14. Schneuing, A. *et al.* Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695* (2022).

15. Qiao, Z., Nie, W., Vahdat, A., Miller III, T. F. & Anandkumar, A. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence* 1–14 (2024).

16. Google DeepMind AlphaFold Team & Isomorphic Labs Team. Performance and structural coverage of the latest, in-development alphafold model (2023). Accessed: 2023-04-16.

17. Krishna, R. *et al.* Generalized biomolecular modeling and design with rosettafold all-atom. *Science* eadl2528 (2024).

18. Evans, R. *et al.* Protein complex prediction with alphafold-multimer. *biorxiv* 2021–10 (2021).

19. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).

20. Chen, J. *et al.* Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions. *arXiv preprint arXiv:2204.00300* (2022).

21. Boyd, N. *et al.* Atom-1: A foundation model for rna structure and function built on chemical mapping data. *bioRxiv* 2023–12 (2023).

22. Celaj, A. *et al.* An rna foundation model enables discovery of disease mechanisms and candidate therapeutics. *bioRxiv* 2023–09 (2023).

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

24. Liu, S. *et al.* Pre-training molecular graph representation with 3D geometry. *International Conference on Learning Representations* (2021).

25. Zaidi, S. *et al.* Pre-training via denoising for molecular property prediction. In *The Eleventh International Conference on Learning Representations* (2022).

26. Atz, K., Grisoni, F. & Schneider, G. Geometric deep learning on molecular representations. *Nature Machine Intelligence* 1–10 (2021).

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

28. Fang, X. *et al.* Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence* **4**, 127–134 (2022).

29. Moesser, M. A. *et al.* Protein-ligand interaction graphs: Learning from ligand-shaped 3d interaction graphs to improve binding affinity prediction. *bioRxiv* 2022–03 (2022).

30. Yan, J. *et al.* Multi-task bioassay pre-training for protein-ligand binding affinity prediction. *arXiv preprint arXiv:2306.04886* (2023).

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

32. Li, S. *et al.* Structure-aware interactive graph neural networks for the prediction of protein-ligand binding affinity. KDD '21 (Association for Computing Machinery, New York, NY, USA, 2021).

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

34. Meller, A. *et al.* Predicting the locations of cryptic pockets from single protein structures using the pocketminer graph neural network. *Nature Communications* (2023).

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

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

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

38. Tsaban, T. *et al.* Harnessing protein folding neural networks for peptide–protein docking. *Nature communications* **13**, 176 (2022).

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

40. Lei, Y. *et al.* A deep-learning framework for multi-level peptide–protein interaction prediction. *Nature communications* **12**, 5465 (2021).

41. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods* **17**, 184–192 (2020).

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

43. Gainza, P. *et al.* De novo design of protein interactions with learned surface fingerprints. *Nature* 1–9 (2023).

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

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

46. Renaud, N. *et al.* Deeprank: a deep learning framework for data mining 3d protein-protein interfaces. *Nature communications* **12**, 7068 (2021).

47. Lam, J. H. *et al.* A deep learning framework to predict binding preference of rna constituents on protein surface. *Nature communications* **10**, 4941 (2019).

48. Xia, Y., Xia, C.-Q., Pan, X. & 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**, e51–e51 (2021).

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

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

51. Sun, L. *et al.* Predicting dynamic cellular protein–rna interactions by deep learning using in vivo rna structures. *Cell research* **31**, 495–516 (2021).

52. Rube, H. T. *et al.* Prediction of protein–ligand binding affinity from sequencing data with interpretable machine learning. *Nature Biotechnology* **40**, 1520–1527 (2022).

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

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

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

56. Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28**, 235–242 (2000).

57. Cheng, F. *et al.* Comprehensive characterization of protein–protein interactions perturbed by disease mutations. *Nature genetics* **53**, 342–353 (2021).

58. Xiong, D. *et al.* A structurally informed human protein–protein interactome reveals proteome-wide perturbations caused by disease mutations. *Nature Biotechnology* 1–15 (2024).

59. Perdigão, N. *et al.* Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences* **112**, 15898–15903 (2015).

60. Barrio-Hernandez, I. *et al.* Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645 (2023).

61. Kulkarni, P. & Uversky, V. N. Intrinsically disordered proteins: the dark horse of the dark proteome. *Proteomics* **18**, 1800061 (2018).

62. Hermosilla, P. *et al.* Intrinsic-extrinsic convolution and pooling for learning on 3d protein structures. *International Conference on Learning Representations* (2021).

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

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

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

66. Batzner, S. *et al.* E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications* **13**, 2453 (2022).

67. Musaelian, A. *et al.* Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications* **14**, 579 (2023).

68. Townshend, R. J. *et al.* Geometric deep learning of rna structure. *Science* **373**, 1047–1051 (2021).

69. Luo, S. *et al.* One transformer can understand both 2d & 3d molecular data. In *The Eleventh International Conference on Learning Representations* (2022).

70. Zaidi, S. *et al.* Pre-training via denoising for molecular property prediction. In *The Eleventh International Conference on Learning Representations* (2023).

71. Zhou, G. *et al.* Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations* (2023).

72. Godwin, J. *et al.* Simple GNN regularisation for 3d molecular property prediction and beyond. In *International Conference on Learning Representations* (2022).

73. Jin, W. *et al.* Dsmbind: Se (3) denoising score matching for unsupervised binding energy prediction and nanobody design. *bioRxiv* 2023–12 (2023).

74. Luo, S. *et al.* One transformer can understand both 2d 3d molecular data (2023). 2210.01765.

75. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic acids research* **49**, D412–D419 (2021).

76. Xia, Y., Zhao, K., Liu, D., Zhou, X. & Zhang, G. Multi-domain and complex protein structure prediction using inter-domain interactions from deep learning. *Communications Biology* **6**, 1221 (2023).

77. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

78. Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* **115**, E3635–E3644 (2018).

79. Pan, J. *et al.* Sparse dictionary learning recovers pleiotropy from human cell fitness screens. *Cell systems* **13**, 286–303 (2022).

80. Adasme, M. F. *et al.* Plip 2021: Expanding the scope of the protein–ligand interaction profiler to dna and rna. *Nucleic acids research* **49**, W530–W534 (2021).

81. Ma, B., Elkayam, T., Wolfson, H. & Nussinov, R. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences* **100**, 5772–5777 (2003).

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

83. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W. (eds.) *Advances in Neural Information Processing Systems*, vol. 34, 29287–29303 (Curran Associates, Inc., 2021).

84. Kaplan, J. *et al.* Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

85. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**, 1257601 (2015).

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

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

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

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

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

91. Kratz, A. *et al.* A multi-scale map of protein assemblies in the dna damage response. *Cell Systems* **14**, 447–463 (2023).

92. Zheng, F. *et al.* Interpretation of cancer mutations using a multiscale map of protein systems. *Science* **374**, eabf3067 (2021).

93. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nature genetics* **50**, 1161–1170 (2018).

94. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with alphamissense. *Science* **381**, eadg7492 (2023).

95. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics* **55**, 1512–1522 (2023).

96. Buniello, A. *et al.* Open targets platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Research* **53**, D1467–D1475 (2025).

97. Szklarczyk, D. *et al.* The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research* **51**, D638–D646 (2023).

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

99. Ochoa, D. *et al.* Open targets platform: supporting systematic drug–target identification and prioritisation. *Nucleic acids research* **49**, D1302–D1310 (2021).

100. Yen, H.-Y. *et al.* Pip2 stabilises active states of gpcrs and enhances the selectivity of g-protein coupling. *Nature* **559**, 423–427 (2018). Author manuscript; available in PMC: 2019 Jan 11. Published in final edited form as: Nature. 2018 Jul 11.

101. Yen, H.-Y. *et al.* Ptdins (4, 5) p2 stabilizes active states of gpcrs and enhances selectivity of g-protein coupling. *Nature* **559**, 423–427 (2018).

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

103. McNally, E. M. β-myosin heavy chain gene mutations in familial hypertrophic cardiomyopathy: The usual suspect? *Circulation Research* **90**, 246–248 (2002). Originally Published 22 February 2002, Free Access.

104. Despond, E. A. & Dawson, J. F. Classifying cardiac actin mutations associated with hypertrophic cardiomyopathy. *Frontiers in Physiology* **9**, 405 (2018).

105. Matsuda, T. *et al.* H-ras isoform mediates protection against pressure overload–induced cardiac dysfunction in part through activation of akt. *Circulation: Heart Failure* **10**, e003658 (2017).

106. Vincent, A., Lang, B. & Kleopa, K. A. Autoimmune channelopathies and related neurological disorders. *Neuron* **52**, 123–138 (2006).

107. Wulff, H. & Zhorov, B. S. K+ channel modulators for the treatment of neurological disorders and autoimmune diseases. *Chemical reviews* **108**, 1744–1773 (2008).

108. Wulff, H., Castle, N. A. & Pardo, L. A. Voltage-gated potassium channels as therapeutic targets. *Nature reviews Drug discovery* **8**, 982–1001 (2009).

109. Yoon, J. H. *et al.* Brain lipidomics: From functional landscape to clinical significance. *Science advances* **8**, eadc9317 (2022).

110. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature reviews genetics* **12**, 56–68 (2011).

111. Judge, S. I. & Bever Jr, C. T. Potassium channel blockers in multiple sclerosis: neuronal kv channels and effects of symptomatic treatment. *Pharmacology & therapeutics* **111**, 224–259 (2006).

112. Craner, M. J. *et al.* Molecular changes in neurons in multiple sclerosis: altered axonal expression of nav1. 2 and nav1. 6 sodium channels and na+/ca2+ exchanger. *Proceedings of the National Academy of Sciences* **101**, 8168–8173 (2004).

113. Parekh, S. *et al.* BCL6 programs lymphoma cells for survival and differentiation through distinct biochemical mechanisms. *Blood, The Journal of the American Society of Hematology* **110**, 2067–2074 (2007).

114. Gutierrez-Camino, A. *et al.* Characterisation of FLT3 alterations in childhood acute lymphoblastic leukaemia. *British Journal of Cancer* **130**, 317–326 (2024).

115. Cairns, R. A. *et al.* IDH2 mutations are frequent in angioimmunoblastic T-cell lymphoma. *Blood, The Journal of the American Society of Hematology* **119**, 1901–1903 (2012).

116. Rodríguez, D. *et al.* Mutations in CHD2 cause defective association with active chromatin in chronic lymphocytic leukemia. *Blood, The Journal of the American Society of Hematology* **126**, 195–202 (2015).

117. Butturini, E., Carcereri de Prati, A. & Mariotto, S. Redox regulation of stat1 and stat3 signaling. *International journal of molecular sciences* **21**, 7034 (2020).

118. Venturutti, L. *et al.* Tbl1xr1 mutations drive extranodal lymphoma by inducing a protumorigenic memory fate. *Cell* **182**, 297–316 (2020).

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

120. Corso, G. *et al.* Deep confident steps to new pockets: Strategies for docking generalization. *ArXiv* arXiv–2402 (2024).

121. Varadi, M. *et al.* Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic acids research* **52**, D368–D375 (2024).

122. Durairaj, J. *et al.* Uncovering new families and folds in the natural protein universe. *Nature* **622**, 646–653 (2023).

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

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

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

126. Bhat, S. *et al.* De novo design of peptide binders to conformationally diverse targets with contrastive language modeling. *bioRxiv* 2023–06 (2023).

127. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* 1–3 (2024).

128. Gane, A. *et al.* Protnlm: Model-based natural language protein annotation (2022).

129. Cerdà-Costa, N. & Xavier Gomis-Rüth, F. Architecture and function of metallopeptidase catalytic domains. *Protein Science* **23**, 123–144 (2014).

130. Morishita, Y. *et al.* Fused radical sam and αkg-hexxh domain proteins contain a distinct structural fold and catalyse cyclophane formation and β-hydroxylation. *Nature Chemistry* **16**, 1882–1893 (2024).

131. Blindauer, C. A. *et al.* Multiple bacteria encode metallothioneins and smta-like zinc fingers. *Molecular microbiology* **45**, 1421–1432 (2002).

132. Park, S. *et al.* Nonlabens marina sp. nov., a novel member of the flavobacteriaceae isolated from the pacific ocean. *Antonie van Leeuwenhoek* **102**, 669–676 (2012).

133. Wu, C.-C., Lin, J. L. & Yuan, H. S. Structures, mechanisms, and functions of his-me finger nucleases. *Trends in Biochemical Sciences* **45**, 935–946 (2020).

134. Zhang, L. *et al.* Structural and functional characterization of deep-sea thermophilic bacteriophage gve2 hnh endonuclease. *Scientific reports* **7**, 42542 (2017).

135. Kuras, R. *et al.* Molecular genetic identification of a pathway for heme binding to cytochrome b 6. *Journal of Biological Chemistry* **272**, 32427–32435 (1997).

136. Ruff, K. M. & Pappu, R. V. Alphafold and implications for intrinsically disordered proteins. *Journal of molecular biology* **433**, 167208 (2021).

137. Arai, M., Sugase, K., Dyson, H. J. & Wright, P. E. Conformational propensities of intrinsically disordered proteins influence the mechanism of binding and folding. *Proceedings of the National Academy of sciences* **112**, 9614–9619 (2015).

138. Hitawala, F. N. & Gray, J. J. What does alphafold3 learn about antigen and nanobody docking, and what remains unsolved? *bioRxiv* 2024–09 (2024).

139. Lotthammer, J. M., Ginell, G. M., Griffith, D., Emenecker, R. J. & Holehouse, A. S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Nature Methods* **21**, 465–476 (2024).

140. Ito, T. *et al.* Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proceedings of the National Academy of Sciences* **97**, 1143–1147 (2000).

141. Huttlin, E. L. *et al.* Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell* **184**, 3022–3040 (2021).

142. Satz, A. L. *et al.* Dna-encoded chemical libraries. *Nature Reviews Methods Primers* **2**, 3 (2022).

143. Söderberg, O. *et al.* Characterizing proteins and their interactions in cells and tissues using the in situ proximity ligation assay. *Methods* **45**, 227–232 (2008).

144. Selvin, P. R. The renaissance of fluorescence resonance energy transfer. *Nature structural biology* **7**, 730–734 (2000).

145. Katsamba, P. S., Park, S. & Laird-Offringa, I. A. Kinetic studies of rna–protein interactions using surface plasmon resonance. *Methods* **26**, 95–104 (2002).

146. Queen, O. *et al.* Procyon: A multimodal foundation model for protein phenotypes. *bioRxiv* 2024–12 (2024).

147. Qin, Y. *et al.* A multi-scale map of cell structure fusing protein images and interactions. *Nature* **600**, 536–542 (2021).

148. Bunne, C. *et al.* How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell* **187**, 7045–7063 (2024).

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

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

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

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

153. Liu, Y. *et al.* Spherical message passing for 3d graph networks. *arXiv preprint arXiv:2102.05013* (2021).

154. Geiger, M. & Smidt, T. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453* (2022).

155. Godwin, J. *et al.* Simple gnn regularisation for 3d molecular property prediction & beyond. *arXiv preprint arXiv:2106.07971* (2021).

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

157. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239* (2020).

158. Liaw, R. *et al.* Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118* (2018).

159. Akiba, T., Sano, S., Yanase, T., Ohta, T. & 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).

160. Li, L. *et al.* A system for massively parallel hyperparameter tuning. *Proceedings of Machine Learning and Systems* **2**, 230–246 (2020).

161. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).

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

163. Biewald, L. Experiment tracking with weights and biases (2020). Software available from wandb.com.

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

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

166. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An open source software for exploring and manipulating networks (2009).

167. Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics* **82**, 949–958 (2008).

168. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction (2020). 1802.03426.

169. Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8 (2015).