Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Accelerated Deep Learning for Structural Protein Docking across the Human Proteome: Cardiovascular Network Translation and Druggability Assessment

VISHWAMBHAR BHANDARE

Vishwambhar.bhandare@partex.ai

Partex.AI India Pvt Ltd. https://orcid.org/0000-0001-8289-1959

Samadhan Khatal
Partex.AI India Pvt Ltd.

Jürgen Scheele
Partex.AI NV

Amit Agarwal
Partex.AI India Pvt. Ltd.

Ketan Sarode
Partex.AI India Pvt. Ltd.

Subhodeep Chatterjee
Partex.AI India Pvt. Ltd.

Urmi Kondhalkar
Partex.AI India Pvt. Ltd. https://orcid.org/0009-0002-1605-0582

Prasanna Joshi
Partex.AI India Pvt. Ltd.

Aman Ahluwalia
Partex.AI India Pvt. Ltd.

Rohit Yadav
Partex.AI India Pvt. Ltd.

Article

Keywords: Novel Protein–protein interactions (PPI), NVIDIA GPU H100, BioNeMo EquiDock, Cardiovascular dis-ease, Genome-wide association studies (GWAS), Druggability assessment

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11761451/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Accelerated Deep Learning for Structural Protein Docking across the Human Proteome: Cardiovascular Disease Network Translation and Druggability Assessment

Vishwambhar Bhandare[a#], Samadhan Khatal[a], Jürgen Scheele[b], Amit Agarwal[a], Ketan Sarode[a], Subhodeep Chatterjee[a], Urmi Kondhalkar[a], Prasanna Joshi[a], Aman Ahluwalia[a], Rohit Yadav[a]

[a.] PARTEX.AI INDIA PVT LTD, 7th Floor, Midas Tower, Hinjawadi, Pune - 411057, Maharashtra, India.
[b.] Partex.AI NV, Frankfurter Strasse 27, Eschborn - 65760, Germany

Corresponding author email: vishwambhar.bhandare@partex.ai

Abstract

Characterizing the protein-protein interactions (PPI's) enable novel therapeutic strategies and defining precise disease mechanisms for human diseases including cardiovascular space. However, vast gap remains between theoretical PPI's search space and physically validated data due to the high cost, low throughput, and high false-positive rates. We present an AI-driven, high-performance computing framework to map the human structural interactome and identify novel therapeutic PPIs for cardiovascular diseases. To overcome historical structure-availability and throughput bottlenecks, we combined proteome-wide AI-predicted structures (Published by Innophore) with NVIDIA's EquiDock, an end-to-end deep learning docking framework. Deployed across 460 NVIDIA's H100 GPUs, this interaction-aware pipeline evaluated a curated matrix of 3.3 million co-localized protein pairs, successfully reducing rigid docking runtimes from hours to seconds. We calculated binding free energies using PRODIGY and integrated these deep learning-derived PPIs with population-scale GWAS datasets, like CARDIoGRAM and Pan-UK Biobank, to map cardiovascular polygenic networks.

The draggability assessment for these novel targets was done using a PocketVec-driven chemogenomic classifier and rDock screening. Cross-referencing our predictions against benchmark interactome repositories (HuRI, BioGRID, STRING, PDBbind) isolated hundreds of thousands of high-confidence, novel physical interactions. Retrospective literature benchmarking confirmed that this computational framework prospectively anticipated subsequent wet-lab discoveries, including TCAP–COLQ and SDHA–IDH2.

Keywords: Novel Protein–protein interactions (PPI); NVIDIA GPU H100; BioNeMo EquiDock; Cardiovascular disease; Genome-wide association studies (GWAS); Druggability assessment

Introduction

Most of the protein functions through interconnected protein-protein interaction (PPI) networks[1]. These physical associations govern nearly every fundamental biological process, maintaining cellular homeostasis and driving disease pathogenesis when dysregulated[2,3]. Consequently, characterizing the human interactome at a structural level is essential for defining precise molecular mechanisms and identifying high-confidence therapeutic targets for drug discovery process[4]. Recently flagship initiative for human proteome project by human proteome organization has been successfully completed for functional annotation of human proteins. This would scaleup multi-omics study of human health and disease biology[5]. However as the number of characterized proteins has increased, the theoretical combinatorial search space of potential interacting protein-protein pairs has exploded into millions of candidates. Despite the necessity of mapping these interactions, a significant gap exists between the theoretically predicted interactome and physically validated data. Traditional experimental protocols, such as yeast two-hybrid (Y2H) screening and affinity purification-mass spectrometry (AP-MS) have inherent limitations owing to high costs, low throughput, and high false-positive rates on a proteome-wide scale[6]. Consequently, major large-scale experimental databases collectively account for less than 100,000 high-confidence binary physical interactions, leaving the vast majority of the human structural interactome entirely unmapped[7,8].

Historically, bridging this gap using computational methods was hindered by a two-fold bottleneck : the scarcity of three-dimensional structural templates and the computational constraints of traditional biomolecular docking algorithms. For decades, a significant fraction of the proteome was inaccessible to structure-based pipelines due to a lack of experimentally derived structures[9,10]. Moreover, classical rigid-body docking suites (such as PIPER, ZDOCK, and HADDOCK) relied on intensive fast Fourier transform (FFT) sampling or multi-stage physics-based refinements requiring minutes to hours per individual protein pair, rendering the generation of proteome-wide interaction matrices computationally unmanageable[8,11,12]. Recent breakthroughs in deep learning have systematically dismantled these structural barriers. The advent of AlphaFold2 (AF2) and the AlphaFold Protein Structure Database has provided high-accuracy, 3D predictions for nearly the entire human proteome[13]. Simultaneously, geometric deep learning and diffusion-based generative frameworks have revolutionized structural docking, predicting highly accurate binding poses in a fraction of the time demanded by legacy physics engines[14,15]. However, large-scale interaction mapping has historically been restricted not only by algorithm design, but also by severe hardware capacity and expansion challenges.

To scale these geometric architectures to proteome-wide screening levels, we leveraged the NVIDIA BioNeMo framework, an enterprise-grade GPU-accelerated generative AI ecosystem purpose-built for biomolecular modelling of protein-protein interactions. BioNeMo integrates state-of-the-art structural and geometric deep learning model EquiDock within a single scalable environment[16,17]. By utilizing specialized TensorRT inference engines and custom CUDA kernels optimized for macro-molecular structures, the framework achieves sub-second processing speeds per protein pair[18–20]. This advanced software layer enables multi-node GPU scaling, effectively compressing a multi-million-pair computational search space into a high-performance, high-fidelity discovery workflow.

Mapping the structural interactome of the cardiovascular disorder (CVD) space represents an urgent therapeutic need considering the yearly global death burden[21–23]. Traditional pharmacological strategies for CVD have focused heavily on a narrow, over-exploited class of surface receptors and enzymes. However, many critical cardiac pathways are controlled by structural proteins or transcription factors that lack these neat binding pockets and have historically been deemed "undruggable." This conventional, single-target approach frequently induces off-target toxicities or triggers compensatory systemic pathways that limit long-term efficacy[24]. Target selection within novel cardiovascular space PPI networks allows for the design of highly specific allosteric modulators capable of disrupting or stabilizing precise macromolecular signaling complexes without broadly shutting down absolute protein function.

Coronary artery disease (CAD) is specific type of cardiovascular disease (CVD) a complex polygenic disorder influenced by numerous interacting genetic and environmental factors. Genome-wide association studies (GWAS) have identified more than 160 loci associated with CAD susceptibility; however, many associated variants remain difficult to interpret mechanistically due to weak individual effect sizes, linkage disequilibrium structure, and pathway pleiotropy[25].

Traditional GWAS analyses prioritize single nucleotide variants (SNVs) surpassing stringent genome-wide significance thresholds. While effective for identifying major loci, this approach may fail to capture distributed biological programs involving multiple weak-effect variants acting within coordinated pathways. Prior work using VEGAS2 and MAGMA gene-based association analyses demonstrated that pathway aggregation reveals additional CAD-associated biology involving coagulation, inflammation, neuronal aging, and wound healing pathways beyond canonical lipid metabolism programs[26].

Importantly, many genes implicated in CAD pathway analyses contain variants that individually fail to reach genome-wide significance yet collectively contribute to disease susceptibility. This observation suggests that CAD risk may emerge through coordinated interaction programs rather than isolated molecular events. Deep learning-derived protein interaction manifolds capture these latent biological structures from sequence, structural, or multimodal biological data, providing complementary systems-biology priors for interpreting polygenic disease architectures[27,28].

Protein-protein interaction (PPI) networks provide a systems-level framework for organizing molecular function. Recent advances in machine learning and representation learning have enabled the construction of AI-derived protein interaction manifolds capable of capturing latent biological structure from sequence, structural, or multimodal biological data. Such learned representations may provide complementary priors for interpreting polygenic disease architectures.

Here, we investigate whether AI-derived protein interaction networks recover biologically meaningful CAD-associated interaction programs from independent UK Biobank and CARDIoGRAM genetic studies. Rather than evaluating direct biochemical binding validation, we assess whether learned interaction priors enrich for cardiovascular disease-associated biological organization, including genes and pathways supported through independent population-scale genetics analyses.

Further, identifying a novel cardiovascular PPI pair, however, is only a foundational step in the discovery timeline. Translating a raw interactome map into therapeutic actions requires an automated method to determine if the individual interfaces or adjacent domains possess targetable pockets capable of binding small molecules. To bridge this critical gap between network discovery and small-molecule chemical validation, we proposed to engineer a PocketVec-driven high-throughput druggability characterization pipeline (https://github.com/sbnb-irb/pocketvec).

Methodology

AI-Driven Molecular Docking for proteome wide PPI's

Human protein sequences and functional metadata were retrieved from the UniProtKB[29] to establish a high-confidence reference baseline for large-scale proteomic discovery[31]. The Innophore has recently published three-dimensional structures of all the annotated proteins (~20k) from UniProtKB[30]. where, high confidence *in silico* structural models were predicted using the OpenFold[31], AlphaFold[32] and ESM-Fold[33]. To purge poorly resolved or unstructured regions, all *in silico* models were processed through the AlphaCutter refinement[34]. For the remaining uncharacterized fraction. This dataset was then subjected to a two-stage filtration process, where the first phase consisted of a sequence length filter which eliminated any protein possessing a sequence length exceeding 1000 residues, followed by structural resolution and topological rigidity filter enforced by systematically discarding any model possessing unresolved segments or disordered loop regions exceeding 7 consecutive amino acids.

Classical physics-based protein-protein docking techniques are highly computationally demanding, requiring ~ 35 minutes to 5 hours per protein complex depending on system size and sampling depth. Consequently, executing classical docking algorithms for a proteome-wide screen containing millions of potentials pairwise protein combinations is entirely infeasible[35]. To overcome this computational bottleneck, EquiDock, an AI-driven rigid-body protein–protein docking framework, was implemented. EquiDock uses equivariant graph neural networks (EGNNs) that work directly on SE (3)-invariant three-dimensional representations of protein structures, which predict the optimal rigid-body transformation placing two protein surfaces in a geometrically and chemically complementary binding pose. Additionally, EquiDock performs docking inference in a single neural network forward pass, omitting the requirement for repetitive energy minimization or extensive conformational sampling loops[36].

NVIDIA 460 H100 GPUs with 36TB VRAM were utilized to parallelize and accelerate the large-scale docking operations[37], where docking runtime was lowered from the minutes or hours required by the classical docking engines to only a few seconds per complex. EquiDock predicted the docked complex shape for every protein pair and the representative docked complex was then chosen from the highest-confidence predicted binding pose for each analyzed pair for subsequent validation **(Figure 1)**.

When expanded to an unconstrained all-by-all matrix, the initial proteome establishes a baseline theoretical search space of 115M unique protein-protein pairs. Applying strict subcellular compartmentalization

constraints onto this space then eliminated geographically impossible combinations, hence refining operational search space down to 27.5M unique, co-localized pairs.

In order to evaluate the resulting docking structural models, PRODIGY (PROtein binDIng enerGY prediction) was utilized to calculate the binding affinity (ΔG, kcal/mol) and dissociation constant (K_d) of the protein–protein complexes, based on quantity and type of intermolecular contacts at the binding interface[38] while also fully characterizing critical intermolecular contacts formed at each predicted PPI interface for all the docked complexes.

Using criteria derived from binding affinities of known protein–protein complexes documented in PDB, all PPIs were then categorized into four groups (*very strong, strong, moderate,* and *weak* binders) based on the projected binding free energies[39].

Filtering of high confidence PPI pairs

To minimize false-positives, further criteria such as missing residue patches, sequence length, subcellular localization, and binding free energy were applied as sequential filters.

A hybrid structure retrieval approach was taken to construct a comprehensive proteome-wide structural dataset. High-resolution 3D structures were derived directly from the Protein Data Bank (PDB) for proteins with existing crystallographic or cryo-EM validation. Binding affinity prediction was done on all the known experimental Protein-Protein complexes. These docked structures were expanded combinatorially and filtered purely based on a thermodynamic threshold enforcing a binding energy threshold of < 100 kcal/mol. This parallel study resulted in an expanded validation matrix, which was evaluated independently against reference databases to establish baseline sensitivity. The second phase involved pair-level filtering of an unlocalized all-by-all matrix generated from the high-resolution monomers. First, a structural pair filter executed shape-complementarity scans by rejecting pairs unfavorable binding affinity values either due to with steric clashes or mismatched surface charge topologies. This was then followed by subcellular localization filters wherein pairs restricted to mutually exclusive cellular environments were discarded. This sequence resulted in a co-localized structural matrix containing 3.3M unique pairs with high confidence optimized for downstream validation.

Generation of Unbiased Protein Pairs

To ensure a completely unbiased, proteome-wide screening architecture free from *a priori* selection bias or premature candidate rejection, every possible pairwise combination was computationally generated from the curated structural dataset. For a dataset containing (n) structurally valid proteins, the unconstrained, non-directional combinatorial search space is mathematically defined by the following combinatorial formula:

$$Totalpairs = \frac{[n(n-1)]}{2} \qquad \text{--------(1)}$$

Here, (n) represents the total number of unique structural proteins, while (n - 1) prevents the uninformative pairing of a protein with itself. The product is divided by 2 because the spatial and biophysical interactions are structurally unordered i.e. Protein A + Protein B is physically and biochemically equal to Protein B + Protein A.

Benchmark Database Acquisition and Curation

To evaluate the overlap with known PPI interactions and identify novel interactions, the predicted PPIs were cross-referenced against several well-established experimental interaction datasets such as HURI, STRING, BioGRID, PDBbind etc. Each database was pre-filtered and standardized against the human Uni-ProtKB ID's reference cohort to establish a clean benchmark framework.

The Human Reference Interactome (HuRI) is a systematic binary protein interaction map covering about 90% of the protein-coding genome created using yeast two-hybrid (Y2H) screening[40]. Raw entries were mapped to the database, dropping obsolete or non-human IDs which were then restricted to the 16k UniProtKB cohort followed by deduplication and refinement at strict thresholds. BioGRID (v4.4) is a curated archive of genetic and protein interactions from literature curation and large-scale interaction screens[41]. Raw inputs were cross-referenced to the 16k human cohort, directional and experimental duplicates condensed refined to high-confidence structural rows in secondary iterations. STRING (v12.0) is a database of known and anticipated protein connections including text mining, co-expression, co-occurrence, and direct experimental evidence[42]. Raw entries were evaluated, filtered and then mapped to standard UniProtKB accessions followed by restricting to the 16k cohort. PDBbind is a curated library of experimentally observed binding affinities for biomolecular complexes with known crystal structures[43]. The binding free energies computed by PRODIGY were compared directly to empirically determined values for structurally similar complexes, which maintained a stable reference of physical complexes across data cleaning steps. Predicted PPIs not captured by any of these four databases were classified as potentially novel.

Multi-Phase Sensitivity Benchmarking and Novelty Isolation

Cross-referencing the computational candidates against these repositories was structured into a two-phase sensitivity framework designed to separate baseline accessions from refined physical structural alignment. The unlocalized structural matrix and the co-localized structural matrix were first mapped directly against the cleaned baseline reference databases to establish the broad capacity of the pipeline to recover documented accessions. Then pipeline's recall precision was re-evaluated exclusively against the isolated high-resolution structural subsets from the databases to mathematically confirm that the pipeline's geometric filters align with known structural biology. These external networks were later merged into two unified reference matrices: a consolidated master reference benchmark of unique human interactions, and a structurally filtered sub-benchmark of unique interactions. Discovery novelty was quantified using a boolean subtraction operation where any pair within our finalized Co-Localized Structural Matrix that returned a non-match against the consolidated master reference benchmark was classified as a completely uncharacterized, novel protein-protein interaction (PPI).

Multi-tier target refinement and dataset partitioning

The computational workflow was split into three distinct validation paths to narrow down a massive pool of millions of potential pairs into high-confidence discoveries. First, an initial energy-based filter was applied to all 115M entries and protein pairs with weak, unfavorable binding thermodynamic properties were discarded, while only the "very strong," "strong," and "medium" binding categories were retained yielding a cohort of 113M entries.

General interactome control pipeline (top 1% validation)

A general control pool was isolated to evaluate how well the NVIDIA BioNeMo EquiDock framework predicts protein interactions. The 113M matrix was subjected to sequential filtering criteria to retain high-confidence interactions and the top-scoring 1% of this dataset was then extracted for further analysis. To validate these computational topologies against native cell properties without introducing circular bias, a *Frequency-Dependent Extraction Protocol* was applied. The top 20 proteins exhibiting the highest pairing frequency for either node position (P1 or P2) within the 1% matrix were extracted and their individual compartmental locations were verified through independent UniProtKB localization metadata.

High-priority cardiovascular target pipeline

Having established that the baseline model is highly accurate, we narrowed our focus to cardiovascular health and assembled a targeted cohort of potential cardiovascular protein pairs. Pairs were first filtered based on strict sequence length limits and a maximum missing loop length. These were then cross-referenced against BioGRID, HuRI, PDBbind, and STRING, which separated the cohort into known and novel pairs. The novel pairs underwent secondary validation and were mapped back against the 115M internal binding energy database. The top 1% highest-scoring binding pairs were selected for deep functional exploration.

Orthogonal pathway network scoring architecture

The structurally valid cardiovascular pairs were mapped against pathway databases to evaluate functional alignment. These pathway-aligned targets were scored via a Composite Pathway Score (CS), where Depth Score evaluates the overall strength of the evidence normalized interaction counts and expert database weights while Evidence Score reflects consistency across different repositories, calculated as an integer score from 1 to 6 based on the matches across six different functional database files.

$$CS = DepthScore\left(\frac{EvidenceScore}{6}\right) \text{----------(2)}$$

Tissue-specific Co-expression and Text-mining Expansion

To further expand the pipeline, a tissue-specific discovery pipeline was employed. Since structural fit alone does not guarantee that two proteins cross paths *in vivo,* a pool of potentially co-expressed protein pairs native to cardiovascular tissue types was generated across unique human UniProtKB IDs. Structural filters were applied to a larger cohort of tissue-specific pairs, which evaluated sequence length and missing loop filters, ultimately yielding unlocalized, structurally valid pairs. Cross-referencing the co-expressed pool with the structurally valid matrix isolated tissue-localized structural pairs. These pairs were then mapped against primary and secondary reference checks to identify novel candidates. For these novel entries, an automated literature verification protocol was written and executed using Perplexity AI to evaluate text-mining confidence classifications. A literature-based pathway constraint was then applied to this dataset, where pairs were retained only if at least one interacting protein possessed a documented role in functional biological ontologies, followed by a final text-mining validation pass against PubMed records.

Coronary Artery Disease system study design and integration pipeline

To systematically resolve the network-level architecture of coronary artery disease (CAD), a comprehensive computational biology pipeline was engineered to integrate deep learning-based interactome predictions with population-scale genetic datasets. The experimental design continuously processes four parallel data channels consisting of *in silico* protein-protein interaction (PPI) networks derived from a sequence-based deep learning framework, CAD-associated high-risk genes curated from the CARDIoGRAM and Pan-UK Biobank cohorts, functional biological annotations across pathway systems and statistical network overrepresentation metrics benchmarked against background models. Detailed protocol for the pipeline for coronary artery disease study design and integration pipeline has been explained in the supplementary file.

As represented in the study workflow **(Figure 2)**, candidate selection transitions through an upstream generation and processing node before split-stream evaluation. Following initial PPI mapping and confidence assignment across target CAD loci, the pipeline bifurcates into an empirical Validation Branch, focusing on population Biobank overlap validation, text-mining, and structural network mapping and a mathematical Statistical Enrichment Branch, evaluating topological pathway clustering significance against stochastic network controls.

End to end druggability assessment using PocketVec

A fully operational, automated end-to-end pipeline we designed, developed and deployed to take an arbitrary list of UniProtKB identifiers as input and output a composite druggability score for each target. This scoring system integrated molecular docking energetics (calculated via the rDock engine) with biophysical cavity geometry. To scale the screening pipeline to handle proteome-wide targets, execution was distributed across a high-performance computing (HPC) cluster managed by a SLURM workload scheduler[44]. The workload was structured as a decoupled, parallelized process where individual protein processing pipelines executed independently, allowing near-linear scaling across cluster nodes. On a benchmark 120-core allocation, the pipeline demonstrates a throughput of approximately 7,300 protein targets per 48 hours **(Figure 3)**.

PocketVec (https://github.com/sbnb-irb/pocketvec) is an automated, end-to-end targetability engine that evaluates the druggability of the interactome by generating highly specialized, fixed-length numerical descriptors of protein binding sites to produce a composite druggability score. The methodology utilizes an inverse virtual screening strategy based on chemogenomic principles and integrates molecular docking energetics via the rDock engine[45] with biophysical cavity geometry. By capturing the unique, functional binding fingerprint of a cavity across a curated panel of lead-like chemical probes, PocketVec maps the pocket's true chemical preferences[46]. At a proteome-wide scale, this approach uncovers cryptic, deeply hidden pocket similarities between completely unrelated protein families that share absolutely no sequence or structural homology. It can determine how effectively the PPI's underlying chemical environment can be modulated by a small-molecule drug, unlocking entirely new classes of targetable targets previously deemed "undruggable." The detailed methodology for this PocketVec pipeline is summarized in the Supplementary File.

Results

The interactome pipeline began with a baseline proteomic inventory of 20,505 human UniProtKB IDs. To maximize reliability of data, an initial data-cleaning filter was applied which excluded unreviewed transcripts, fragment records, and ambiguous sequence variants, reducing the data pool to 17,112 IDs. This was filtered further to retain only verified human reference proteins, yielding a reference cohort of 16,164 unique human UniProtKB IDs. An unbiased, all-by-all combinatorial pairing of the baseline proteome gives an unconstrained theoretical search space of 115M unique protein pairs. Approximately 1,700 UniProtKB IDs which lacked verified compartment annotations were excluded. Applying this biological localization restriction removes spatial background noise, dropping the search space down to a localized baseline matrix of 27.5M unique pairs **(Table 1)**.

Using a hybrid approach a structural library of 16,242 individual entries was assembled to transition to physical molecular modeling. To ensure geometric compatibility and optimize downstream docking processes, this library was subjected to sequential filtering – first to eliminate sequences exceeding 1000 amino acids in length and then to discard models exhibiting highly disordered segments or missing loop regions greater than 7 consecutive amino acids, resulting in a final 6005 high-confidence entries.

When mapped and filtered at the pair level for structural compatibility, these models yielded two candidate spaces (i) an Unlocalized Structural Matrix containing 15.5M unique pairs, and (ii) a dual-constrained Co-Localized Structural Matrix containing 3.3M unique candidate pairs **(Figure 1; Table 1)**.

Mapping of predicted PPIs with known interactions networks

To establish a baseline for validation, four external reference databases (BioGRID, HuRI, PDBbind, and STRING) were acquired and subjected to the identical 16,164 human UniProtKB https://doi.org/10.1038/s41597-024-03403-z filter to discard outdated accessions, non-human matches, and multi-source redundancy. From BioGRID an unfiltered pool of 321,476 raw records was cleaned down to 302,747 human-mapped entries. Removal of multi-source publication duplicates yielded 86,131 unique verification pairs. In case of HuRI, an initial pool of 1,280,687 records was cleaned to 54,506 pairs derived from 9,032 unique IDs. For STRING, the initial 1,182,858 text-mined records after filtering yielded 467,412 unique clean pairs. PDBbind retained all entries throughout data cleaning, preserving 415 highly verified crystallographic complex pairs. De-duplicating these into a singular, unified repository resolved dataset overlapping to ultimately establish a consolidated reference benchmark of 529,492 unique observed interactions, alongside a high-resolution structurally compliant sub-benchmark of 91,883 unique interactions **(Table 2)**.

To validate the structural and biological accuracy of the pipeline, sensitivity mapping was done across two distinct phases to separate baseline accession from refined physical structural alignment.

Mapping our candidate spaces directly against the complete, human-filtered reference networks provided an initial verification of pipeline alignment. When evaluated against the unlocalized 15.5M structural matrix, the pipeline demonstrated stable capture rates across all repositories, recovering 12.60% of BioGRID, 15.98% of HuRI, 15.46% of PDBbind, and 15.14% of STRING interactions.

When biological co-localization was enforced within the 3.3M matrix, low-throughput physical complexes within PDBbind remained unchanged. However, high-throughput screening networks showed steep decline, with HuRI capture dropping to 6.16% (a loss of 61.4% of its structural matches), highlighting the pipeline's ability to differentiate between unlocalized *in vitro* binding ability from actual native physiological co-localization.
When the spatial boundaries were not considered, initial mapping against the unconstrained 15.5 million structural matrix acted as a positive control. The pipeline achieved uniform sensitivity, hence mathematically confirming that the sequence and structural loop filters are strongly aligned with the geometry of known interactomes. However, applying the cellular requirement of shared subcellular localization within the 3.3M matrix had a dramatic effect, with physical complexes frozen directly from cells in PDBbind were highly resilient, high-throughput assay datasets showed steep declines **(Table 3)**.

The multi-million pair matrices were cross-referenced directly against the consolidated master reference benchmarks with the goal to establish a non-redundant global validation baseline. Comparing the unconstrained structural matrix against the structurally filtered sub-benchmark achieved a high database recall rate, confirming that over 86% of historically validated, structurally viable human interactions are successfully preserved within our model. More importantly, when applying the constraint of shared subcellular localization within our 3.3 million candidate matrix, the captured benchmark pairs overlapped resulting in a localized database recall of 45.76%, indicating that 54.24% of structurally viable **(Table 4)**, known interactions involve proteins that are physically segregated in separate cellular compartments under native physiological conditions. By filtering out these non-physiological associations, the pipeline purifies the candidate space, more than doubling the precision density of known interactions within our operational computational pool.

Unbiased Proteome-Wide Discovery of Novel PPIs

The primary objective of this exhaustive filtering and pairing approach was to systematically map the uncharacterized human interactome. By subtracting the compiled observational benchmark sets from our finalized co-localized matrix, we isolated the unmapped space of the human proteome.
Out of the 3,314,683 highly vetted, co-localized structural pairs evaluated, only 42,051 pairs could be accounted for in existing scientific literature and public repositories. Consequently, the pipeline discovered 3,272,632 completely novel, physically plausible, and co-localized human protein-protein interactions (**Table 4)**. Since millions of these pairs have successfully passed both the strict high-resolution structural integrity checks and physical cell co-localization constraints, they represent highly prioritized candidates for downstream high-throughput structural modeling or targeted wet-lab biochemical validation.

Filtering the unconstrained 115M theoretical global matrix to retain only highly favorable binding potentials adjusted the operational search space to 113M pairs creating a mathematically clean and biophysically sound environment. This was then cross-referenced directly against two standard interaction repositories (STRING and BioGRID), narrowing the novel PPI pair to the top 10% highest-scoring entries of 317,962 unique pairs. Further processing was done to extract the top 20 hub interactors, which occurred most frequently for either Protein 1 and/or Protein 2. These were then validated by successfully matching

individual UniProtKB compartmental profiles. Because both nodes in these high-frequency pairs shared matching subcellular addresses, this control group confirmed that the pipeline's unconstrained structural docking scoring strongly aligned with native intracellular compartments.

Cardiovascular space PPI mapping

Applying strict sequence length and missing loop filters to the initial 85,000 cardiovascular candidate pairs successfully removed structural artifacts. Our first-pass database cross-referencing against the four standard databases isolated 212 known pairs, with a significant majority originating from STRING records. The remaining pairs were classified as provisionally novel. The second-pass validation against the expanded V2 database removed additional hidden STRING-predicted entries and found no matches within BioGRID, leaving a final pool of 10,641 clean, novel cardiovascular protein pairs. Mapping this clean pool against the 115M energetic database confirmed that 9,624 pairs satisfied all biophysical stability constraints. Extracting the top 10% highest-scoring entries delivered 962 prioritized novel cardiovascular target pairs (**Table 5**). A list of these high confidence novel PPIs is included in the **Supplementary Table S1**. To verify the functional relevance of the 9,624 validated cardiovascular pairs, they were mapped against public pathway networks isolating 3,861 pairs that shared at least one functional biological pathway. Calculating the Composite Pathway Score (CS) across these pairs provided a balanced measure of evidence depth and source consistency. The calculated CS values were correlated directly with the structural docking free energies (ΔG) in order to evaluate the mathematical relationship between physical structure and network biology, which revealed a statistical correlation of -0.05, confirming that the binding free energy values and pathway-alignment scores represent completely orthogonal, non-redundant layers of evidence. Furthermore, this independent relationship ensures that candidate pairs excelling in both dimensions satisfy both structural compatibility and functional biological necessity, confirming these 962 targets as high-confidence candidates for downstream experimental validation.

Shifting the pipeline from simple compartmental boundaries to strict tissue co-expression profiles unlocked a broader discovery space. Filtering the initial ~49.9M tissue-mapped pairs through sequence length and loop stability filters yielded 15.1M unlocalized, structurally valid pairs. Cross-referencing this structurally sound matrix against the 14.8M tissue co-expressed pool produced tissue-localized structural pairs **(Table 6)**. Our first-pass reference check matched 4,237 known pairs and isolated novel entries. Our second-pass validation successfully flagged previously documented interactions, leaving a final pool of 1,076,657 unique, tissue-specific novel protein pairs. To evaluate the textual validity of this large pool, manual literature review using Perplexity AI was performed across a random sample of 664 entries to classify text-mining confidence layers.

Finally, applying strict pathway ontology filters to the 1,076,657 novel pairs revealed that 41.4% of the novel dataset contained at least one interactor with a verified role in functional biological pathways. Cross-referencing these pathway-mapped pairs against published literature indices identified known mentions, confirming a final discovered pool of 436,841 completely novel, tissue-co-expressed, and pathway-vetted cardiovascular interaction targets.

Construction of the Prioritized CAD Interactome Architecture

The sequence-based deep learning framework (DL-PPI) was applied to execute proteome-wide link predictions, mapping raw primary amino acid sequences onto an undirected interaction graph G= (V, E). By coupling these predictive physical interaction likelihoods with large-scale population genetics metadata collection containing 120,575 individuals from the CARDIoGRAM Exome Cohort and 442,574 individuals from the Pan-UK Biobank cohort, the composite scoring engine effectively elevated sub-threshold polygenic risk signals. A high-confidence target validation framework isolated a core module of 20 driver genes consistently prioritized across both the VEGAS2 and MAGMA analytical pipelines from aggregated genomic profiles as listed in **(Table 7)**.

To map the physical landscape of the predicted networks prior to functional validation, a symmetrical composite-score matrix was constructed, displaying the edge strengths across the 23 prioritized genes comprising the top 50 candidate pairs **(Figure 4)**. The matrix resolved a highly polarized structural topology, separating transient signaling modules from dense, permanently localized multi-subunit complexes.

The dominant feature of this structural interactome is a massive, highly interconnected mitochondrial bioenergetics hub, within which the NADH dehydrogenase iron-sulfur protein 1 **(*NDUFS1*)** acts as the master structural organizer, displaying the highest composite edge scores (exceeding 0.5) across almost all other mitochondrial elements in the dataset. This includes highly coordinated structural linkages to other complex I subunits **(*NDUFA3*, *NDUFC1*, *NDUFA8*)**, complex IV elements **(*COX5A*, *COX5B*)**, and ATP synthase components **(*MT-ATP6*, *ATP5MC3*)**. The high density of these intense reciprocal scores indicates that the deep learning model successfully recognized the tight physical packing and geometric proximity of these proteins within the inner mitochondrial membrane. In contrast, the matrix isolated a distinct, completely self-contained structural module. The fibroblast growth factors ***FGF18*** and ***FGF23*** exhibited a prominent mutually predicted interaction (composite score ≈0.4) yet remained completely decoupled from the rest of the 21-gene network. This isolation perfectly captures the unique nature of this extracellular signaling axis, confirming that the structural modeling pipeline can successfully differentiate highly specific ligand-receptor interactions from broad, macromolecular metabolic complexes.

To filter out computational noise and validate the biological relevance of the top 50 predicted candidate pairs, their functional annotation depth was profiled across six orthogonal data channels **(Figure 5)**. The multi-evidence pathway overlap matrix revealed distinct physiological clustering regimes based on relative Z-scored pathway co-occurrence. Specifically, the upper and central bands of the interactome, predominantly dominated by ***NDUFS1*** and ***SDHA*** partner configurations (e.g., ***NDUFA3—NDUFS1***, ***NDUFC1—NDUFS1***, ***COX5A—NDUFS1***) demonstrated consistent, robust over representation (Z ≈ 0 to 2) across WikiPathways, GO-BP, and GO-MF, confirming their physical coordination inside metabolic electron transport machinery. In contrast, the isolated extracellular signaling axis ***FGF18—FGF23*** exhibited a hyper-enriched anomaly restricted entirely to Reactome (Z ≈ 6) and WikiPathways (Z ≈ 4). Conversely, it displayed a sharp negative validation signal in cellular component databases (GO-CC, Z≤ -4), distinguishing it from dense, localized inner mitochondrial membrane structural networks.

To demonstrate that the deep learning model successfully captured coordinated biological programs rather than stochastic structural overlaps, we conducted functional over representation testing within the top 10% ranked PPI pairs **(Figure 6)**. The statistical baseline (Odds Ratio = 1.0) was heavily exceeded across two primary physiological regimes, starting with a pronounced signal in cytoskeletal assembly and muscle morphogenesis. Within this structural regime, pathways displayed the largest effect sizes across the prioritized decile, where Actomyosin Structure Organization and Myofibril Assembly achieved peak Odds Ratios between 65 and 70. Simultaneously, Supramolecular Fiber Organization was identified as the most statistically significant functional cluster in the entire analysis ($-log_{10}$FDR) > 52.5). Consistent, high significance over representation was mapped across these vital metabolic programs of bioenergetic transduction pathways, which included Oxidative Phosphorylation, Aerobic Respiration, and Cellular Respiration. Furthermore, this bioenergetic signature extended to multiple Monoatomic Ion Transmembrane Transport pathways, with the entire metabolic cohort maintaining robust statistical significance bounds ($-log_{10}$FDR)≈ 32.5 to 45.0).

To bridge the structural interactome directly with phenotypic CAD pathology, the top-ranked decile cohort was evaluated against canonical KEGG and Reactome biological pathway repositories. As illustrated in **Figure 7**, the prioritized KEGG network demonstrated highly focused disease over representation. Dilated Cardiomyopathy (DCM) emerged as the dominant functional driver, displaying maximum structural effect size (Odds Ratio ≈32) and peak statistical significance ($-log_{10}$FDR)>4.5). Strong enrichment was observed for Arachidonic Acid Metabolism and Arg/Pro Metabolism (Odds Ratios≈18), mapping links to plaque stability and vascular inflammation, alongside adjacent structural validation in the MAPK Signaling Pathway and the Regulation of Actin Cytoskeleton. Granular molecular mapping via the Reactome database **(Figure 8)** confirmed a major metabolic convergence where the Aerobic Respiration and Respiratory Electron Transport pathways dominated the architecture, demonstrating an exceptional statistical significance score (log_{10}FDR > 80) and a substantial effect size (Odds Ratio ≈ 20.3). This metabolic backbone was heavily supported by the over representation of Mitochondrial Protein Degradation (Odds Ratio ≈ 13.7) and Fatty Acid Metabolism (Odds Ratio ≈ 9.0). Downstream cellular execution points were anchored through significant over representation in canonical Muscle Contraction and Cardiac Conduction cascades, confirming that the computationally predicted CAD interactome successfully reflects clinical tissue-level homeostatic systems.

Druggability assessment on identified novel cardiovascular targets

To scale the screening pipeline to handle proteome-wide targets, execution was distributed across a high-performance computing (HPC) cluster managed by a SLURM workload scheduler. Moving the PocketVec pipeline from the local commodity workstation to the SLURM-orchestrated 120-core HPC cluster significantly improved execution speeds and processing efficiency. While the prototype workstation (24 cores) successfully completed 1,460 proteins in 48 hours, the production HPC cluster allocation scaled throughput near-linearly to process approximately 7,300 protein targets within the same 48-hour window. The pipeline's sequential stages were optimized according to their computational and resource complexities to mitigate the rate-limiting performance bottlenecks.

The analytical architecture was validated through an empirical screen consisting of 1,839 proteins. The scoring heads successfully ranked targets into discrete high-, medium-, and low-resolution programmatic druggability tiers based on their thermodynamic compatibility and spatial profiles. Favorable Docking refers to probes that achieve a docking energy score of < -5.0, Strong Docking refers to probes that achieve a docking energy score of < -9.0, indicating high-affinity interactions likely to result in biological activity, Optimal Cavity Range refers to cavities with volumes between 200 and 1000 $Å^3$, which are considered ideal to accommodate typical drug-like small molecules and outliers and probes with docking scores > 0, identify unfavorable binding environments where molecular repulsion or poor fit occurs. The final 0–100 druggability score shows a direct correlation with mean probe scores, where high-scoring targets (scores > 75) generally exhibit mean probe scores in the range of -8.0 to -13.0, paired with minimal outliers. Conversely, scores decline precipitously as the mean docking energy increases or the count of outliers rises, representing sites that are either chemically incompatible or structurally undersized.

Proteins categorized within the "High" tier demonstrated exceptionally robust probe interaction profiles (refer to **Supplementary Table S2)**. These targets are characterized by a significant density of favorable and strong docking scores, often retaining excellent scoring profiles despite cavity volumes that exceed the traditional optimal range. Mean probe scores are universally sturdy, typically occurring below -8.0 and frequently dropping below -13.0 for top-ranked targets. Favorable docking counts are almost universally above 100 of 128 probes, while strong docking counts usually exceed 40 per protein, with top candidates showing greater than 100 high-affinity hits. Outlier counts are minimal or entirely absent, indicating highly compatible electrostatic and steric binding environments across most of the active site grid surface.

Medium-tier proteins show moderate small-molecule binding potential (refer to **Supplementary Table S3)**. While they exhibit a significant count of favorable docking events, their overall druggability rankings are systematically attenuated by less favorable mean scores, cavity volumes significantly outside the optimal medicinal chemistry range, or the localized presence of binding site outliers that signal steric or electrostatic clash zones. Targets in this category frequently fall outside the optimal cavity volume range. Their structural configurations frequently present either highly dilated fallback cavities exceeding 1,000$Å^3$ or tightly enclosed pockets showing a significant presence of binding site outliers. While moderate binders manage a high density of favorable interactions, their global thermodynamic anchor is weaker compared to the High-tier profile.

Targets assigned to the "Low" tier exhibit poor computational docking performance across the board (refer to **Supplementary Table S4**). These sites are frequently characterized by a complete absence of high-affinity strong docking scores and a high prevalence of severe outliers, contributing to an unfavorable binding environment dominated by molecular repulsion. Top-tier repulsive targets demonstrate near-total probe rejection, generating outliers which skew their mean scores to highly unfavorable positive values. The lowest performing end of the proteomic spectrum features targets which generate zero favorable hits, zero strong hits, and absolute steric rejection across all 128 library probes, cementing their classification as poorly druggable or un-encapsulated surface architectures. At the absolute boundary limit,

targets drop below the physical cavity envelope entirely, recording zero hits and maximum geometric penalties.

Discussion

Although high-throughput computational tools have drastically expanded our structural coverage of the human proteome[47], a persistent challenge in interactome mapping was separating the superficial structural cross-reactivity from true physiological relevance. Just because two proteins have complementary geometric surfaces does not always guarantee that they ever occupy the same spatial or a time-based alignment window within a living cell[48]. With the goal to systematically address this uncertainty, our workflow implemented a dual-track benchmark that was designed to independently evaluate a protein pair's intrinsic structural compatibility away from its native biological environment. By disconnecting the raw geometric binding capacity from strict subcellular localization constraints, this framework isolates the distinct contributions of physical shape versus cellular geography. This further prevents the false-positive inflation commonly seen in unlocalized *in silico* screens, providing a highly reviewed and validated structural foundation essential for identifying actionable, high-confidence therapeutic intervention points in modern systems medicine[4].

Comparative Analysis of Structural Capacity versus Spatial Constraints

To systematically dissect the dynamics between geometric binding capability and physiological feasibility, we mapped high-resolution structural subsets from reference databases independently against two distinct, multi-million-pair computational spaces. This dual-track benchmark was specifically engineered to separate a protein pair's intrinsic structural compatibility and shape complementarity, from the spatial and environmental constraints of its native biological setting.

In doing so, we address the tendency for unconstrained *in silico* docking algorithms to yield false positives by pairing proteins that have highly complementary surfaces but are strictly segregated within the cell. By evaluating physical docking energy and cellular co-localization as completely independent vectors, this framework was able to establish a clear line of distinction between what is biophysically possible in an isolated system and what is physiologically allowed within a living organism. One track operates under ideal biophysical conditions, treating the proteome as a closed, well-mixed system exposing the binding capacity of the structural proteome, mapping out full cross-reactive potential as well as hidden binding abilities built into human protein structures. In parallel, the second track applies strict environmental boundaries based on observational cellular data. A protein pair cannot form a functional, native interaction if its individual subunits are confined in non-communicating compartments, hence filtering the space through the lenses of subcellular compartmentalization, tissue-specific expression profiles and co-expression dynamics. By deliberately separating these two tracks before crossing them into a final matrix, the pipeline isolated false positives while quantifying geometrically perfect pairs (structural compatibility but physiologically impossible). Additionally, it isolated high-confidence targets with overlapping shape and space, while also identifying hidden interfaces which can be exploited to force non-native interactions for therapeutic purposes.

The Structural Baseline (15.5M) Vs Physiological Reality Check (3.3M)

Initial mapping against the unconstrained 15.5 million structural matrix served as a thorough positive control for our geometric filtering parameters. When spatial boundaries were omitted, our pipeline demonstrated exceptional sensitivity, capturing 84.61% of structurally viable BioGRID pairs, 83.30% of HuRI pairs, 86.87% of STRING pairs, and 100.00% of the crystallographic complexes documented in PDBbind. The remarkably high and uniform recall rates across highly divergent databases mathematically validated the foundational architecture of the structural screening pipeline, proving that the strict resolution thresholds strongly align with the structural configurations of known, real-world interactomes.

Introducing the requirement of shared subcellular localization dramatically altered the interaction landscape and revealed a significant disparity between *in vitro* binding capacity and *in vivo* biological reality. While the total number of target baseline pairs within the reference databases remained fixed, their alignment with our refined 3.3 million co-localized matrix dropped substantially: BioGRID retention decreased from 84.61% to 61.44% (a loss of 4,944 validated matches), HuRI retention dropped from 83.30% to 32.35% (a loss of 7,017 validated matches) and STRING retention fell from 86.87% to 47.22% (a loss of 42,927 validated matches). Conversely, PDBbind, consisting of physical complexes frozen and captured together via structural biology techniques, was resilient to spatial filtering, retaining 70.31% of its matches (45 of its 64 filtered pairs). High-throughput assays (such as Yeast Two-Hybrid screening of HuRI) and automated text-mining networks often document interactions under artificial conditions where native cellular architecture is not preserved. By forcing human proteins into proximity within heterologous cells or test tubes, these assays capture interactions that are structurally possible but physiologically impossible. Our localization pipeline successfully isolated and eliminated this non-physiological background noise, providing a highly filtered candidate pool which was optimized for novel physiological discovery.

A key finding of this study was the capacity of learned interaction representations to de-orphan and contextualizes sub-threshold GWAS signals. In standard genetic association pipelines, critical risk-associated genes are often ignored because their underlying variants do not individually achieve genome-wide significance[49]. Our pipeline bypassed this bottleneck by identifying 70 novel interaction candidate pairs (Category B) involving genes independently implicated in CAD-associated pathway analyses. While the individual genomic signals for these Category B genes remained obscured by statistical noise in isolated variant analyses, aggregating them into a physical network structure revealed that they converge into highly coordinated, biologically coherent modules. By using composite scoring strategies (integrating KEGG, Reactome, Wiki Pathways, and GO domains) to rank model confidence, we demonstrated that these sub-threshold genes operate as functional components of larger, coordinated multi-protein complexes.

The functional enrichment of the predicted CAD subnetworks revealed a highly focused convergence of cellular pathology across several physiological programs. While the recovery of coagulation, lipid metabolism, and inflammatory modules aligned with classical CAD pathophysiology, the enrichment of neurovascular and neuronal apoptosis-associated genes supported emerging hypotheses linking progressive vascular dysfunction with shared neurodegenerative pathways. These predicted interaction networks are interpreted as biologically supported, interaction-informed prioritization candidates representing systems-level associations.

The true predictive capacity of a computational pipeline is demonstrated when it's *in silico* outputs anticipate subsequent observational discoveries. This study was executed as an intensive, high-throughput exercise to predict novel PPIs across the human proteome, ultimately generating 3,272,632 novel, co-localized interactions within a rapid 4-day computational window on distributed 460 NVIDIA H100 GPU clusters. Remarkably, during a follow-up analysis conducted several months post-execution, a systematic literature text-mining sweep revealed that 296 of these predicted "novel" interactions had subsequently been independently validated and published in external peer-reviewed research articles. An example of this predictive forecasting is the interaction between Telethonin (TCAP) and the collagen-like tail subunit of acetylcholinesterase (COLQ). While both proteins are integral to neuromuscular and musculoskeletal integrity, their coordinated functional link in muscle maintenance and pathology was highlighted in a 2025 study. TCAP and COLQ were identified as hub regulators in pathways driving muscle development and structural preservation providing strong literature validation for the biological relevance of the TCAP–COLQ interaction axis in muscular dystrophy and neuromuscular health[50]. This retrospective confirmation elevates the pipeline from a descriptive modeling tool to a predictive instrument capable of accelerating discovery timelines by flagging physically viable interactome boundaries months ahead of traditional wet-lab configurations.

High-Performance Scaling and the Chemogenomic Druggability Landscape

Simulating molecular interactions at a proteome-wide scale introduces extreme non-uniform computational demands. The docking stage (utilizing rDock) behaves under a stochastic search complexity which is highly CPU intensive, whereas the downstream Similarity Head scaling quadratically, introduces a severe memory and vector processing bottleneck. By implementing an asynchronous, parallelized architecture under SLURM, we achieved near-linear scaling, processing 7,300 protein targets in the same 48-hour window that previously limited a local workstation to just 1,460. By orchestrating dataflow and hardware resource allocation effectively as demonstrated by this pipeline, scaling structural genomics is no longer limited by algorithm design.

Limitations and Future Trajectories

Despite the predictive power of this deep-learning integration, several limitations must be addressed to guide future research.

1. Hypothesis vs. Physical Proof: The predicted interactions do not constitute direct biochemical validation. At this stage of development, these pairs must be interpreted as biologically plausible, prioritizable hypotheses rather than experimentally proven physical complexes.
2. Database Annotation Dependencies: The sensitivity of our pipeline's downstream enrichment and filtering tools depends heavily on the quality and completeness of existing public registries and PPI training datasets.
3. Lack of Orthogonal Wet-Lab Validation: The current study lacks direct wet-lab confirmation. Future research incorporating functional perturbation assays, CRISPR-based interaction screens, or physical structural docking analyses would substantially strengthen the causal and therapeutic interpretation of these novel CAD candidate targets.

Conclusion

In this study, we established an unbiased, proteome-wide structural screening pipeline that successfully couples rigid-body geometric docking with strict biological spatial constraints. By deploying the AI-driven EquiDock framework accelerated by massive computation by 460 NVIDIA H100 GPUs, we bypassed the severe performance bottlenecks of classical physics-based docking engines, allowing exhaustive exploration of millions of unlocalized human protein pairs. Our pipeline isolated 3,272,632 completely novel, physically plausible, and co-localized human protein-protein interactions. These prioritized candidate pairs represent an optimized baseline for targeted downstream high-throughput structural modeling and direct wet-lab biochemical validation, opening new avenues for therapeutic target discovery.

Beyond baseline interactome mapping, this dual-track architecture proved highly effective when applied to specialized physiological systems. By restricting our search space to cardiovascular health, we successfully isolated a high-priority cohort of 9,624 structurally valid, novel cardiovascular pairs from a broader tissue-specific co-expression matrix of 14.8 million pairs. Coupled with a tissue-specific co-expression filter and real-time semantic literature validation, this multi-tier target refinement pipeline successfully bridged the gap between pure biophysical capability and true *in vivo* cellular geography. To translate these network-level interactions into actionable therapeutic opportunities, we scaled our PocketVec-driven pipeline near-linearly, mapping out a comprehensive, thermodynamic druggability landscape that uncovered cryptic, highly targetable allosteric pockets on previously "undruggable" signaling proteins.

Crucially, the biological relevance of these deep learning–derived interaction manifolds were demonstrated by integrating structural topologies with population-scale genetic architectures from the Pan-UK Biobank and CARDIoGRAM cohorts. By overlaying biobank-scale genome-wide association study (GWAS) summary statistics onto our structural PPI networks, the pipeline successfully captured distributed, weak-effect polygenic risk modules for coronary artery disease that missed traditional single-locus significance thresholds, transforming raw population genetics into mechanistically interpretable physical networks.

While validated here within cardiovascular disease, the underlying computational and biophysical principles of this framework are inherently disease nonspecific. Because the physical rules governing macromolecular interfaces, pocket geometry, and binding energetics are universal, this integrated platform provides a reusable, high-throughput strategy to examine complex polygenic landscapes across diverse clinical domains ranging from identifying allosteric targets in neurodegenerative conditions (such as Alzheimer's and Parkinson's) and oncogenic signaling networks to mapping dysregulated pathways in metabolic and renal disorders. By offering a highly prioritized, druggable map of the human interactome, this work establishes an optimized computational foundation for downstream high-throughput structural modeling and direct wet-lab biochemical validation, opening rapid new pathways for targeted therapeutic discovery in modern systems medicine. To summarize, by unifying enterprise geometric deep learning, distributed HPC infrastructure, and automated chemogenomic pocket vectorization, this framework transforms descriptive interactome mapping into a predictive discovery engine. The platform effectively bridges the gap between proteome-wide structural prediction, chemical targetability, and polygenic

translation, establishing a disease-agnostic blueprint for accelerating network-based drug discovery across complex human phenotypes.

Authors responsibilities

VB: Conception, approach design, execution, PPIs analysis, validation, writing, revision work
SK: Set up and parallel execution of large scale PPI on 460 H100 GPU's to complete work in just 4 days
JS: Draft review and editing, supervision, feedback
RY: Set up and executed PocketVec pipeline for druggability assessment
UK & PJ: Drafting manuscript and figure / table preparation
AA (Aman A)**:** Data analysis and interpretation
AA (Amit A)**:** Conception, analysis, interpretation of data, supervision
KS & SC: Optimization and testing of PPI pipeline end to end flow

Acknowledgments

We gratefully acknowledge Partex.AI Technology Pvt. Ltd. for their continuous encouragement and providing 460 H100 GPU's infrastructure support. Special thanks to Suyash Masugude, DevOps team including Shashank, Nitin Koli, Shrikant Patil, Priti Jadhav, Gaurav Tripathi, and Gunjan Bhardwaj for providing unconditional support all the time for this breakthrough high impact research for novel therapeutic target discovery. Thanks for UK Biobank for sharing the data and Sanyogita Purandare for processing the data and performing analysis. We also thank team NVIDIA Ajay Kumar and Arundhati Banerjee for providing an opportunity to present this breakthrough at NVIDIA GTC 2025.

The writing of this manuscript was assisted by Claude in order to improve language and readability. After using this tool, the authors reviewed and edited the content as needed.

Ethical Considerations

This study did not involve human participants, human data, or human tissue. All protein sequence, structure, and protein pairs data were derived from publicly available sources (UniProtKB, String, HURI, BioGRID, PDBBind etc). This *in-silico* study did not require review or approval by an Institutional Review Board or equivalent ethics committee.

Consent to Participate

Not applicable. This study did not involve human participants.

Consent for Publication

Not applicable. This manuscript does not contain any individual person's data in any form.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by German BMFTR for the CurATime/CurAIknow Grant (03ZU1202NC)

Data Availability Statement

All the AI predicted novel high confidence cardiovascular space associated PPI pairs are being submitted along with manuscript in the supplementary file. Any other relevant raw data would be available for sharing upon request (subject to company policy). In house end to end PPI docking pipeline is proprietary to Partex.AI technology Pvt. Ltd hence cannot be shared / made available for public due to commercial confidentiality.

References

1. Safari-Alighiarloo, N., Taghizadeh, M., Rezaei-Tavirani, M., Goliaei, B. & Peyvandi, A. A. Protein-protein interaction networks (PPI) and complex diseases. *Gastroenterol Hepatol Bed Bench* **7**, 17–31 (2014).

2. Lu, H. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal Transduct Target Ther* **5**, 213 (2020).

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

4. Capriotti, E., Ozturk, K. & Carter, H. Integrating molecular networks with genetic variant interpretation for precision medicine. *Wiley Interdiscip Rev Syst Biol Med* **11, e1443**, (2019).

5. Omenn, G. S. The 2024 Report on the Human Proteome from the HUPO Human Proteome Project. *J Proteome Res* **23**, 5296–5311 (2024).

6. Rao, V. S., Srinivas, K., Sujini, G. N. & Kumar, G. N. S. Protein-Protein Interaction Detection: Methods and Analysis. *Int J Proteomics* 147648 (2014).

7. Garzón, J. I. A computational interactome and functional annotation for the human proteome. *eLife* **5, e18715**,.

8. Eramian, D., Eswar, N., Shen, M. & Sali, A. How well can the accuracy of comparative protein structure models be predicted? *Protein Sci* **17**, 1881–1893 (2008).

9. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The Protein Folding Problem. *Annu Rev Biophys* **37**, 289–316 (2008).

10. Moult, J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr Opin Struct Biol* **15**, 285–289 (2005).

11. Ritchie, D. W., Kozakov, D. & Vajda, S. Accelerating and focusing protein–protein docking correlations using multi-dimensional rotational FFT generating functions. *Bioinformatics* **24**, 1865–1873 (2008).

12. Evans, R. Protein complex prediction with AlphaFold-Multimer. (2021) doi:10.1101/2021.10.04.463034.

13. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nat Commun* **13**, 1265 (2022).

14. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. (2022) doi:10.48550/ARXIV.2210.01776.

15. Stärk, H., Ganea, O.-E., Pattanaik, L., Barzilay, R. & Jaakkola, T. EquiBind: Geometric Deep Learning for Drug Binding Structure Prediction. (2022) doi:10.48550/ARXIV.2202.05146.

16. Overview — NVIDIA NIM for OpenFold2 (https://docs.nvidia.com/nim/bionemo/openfold2/latest/overview.html) .

17. EquiDock — NVIDIA BioNeMo Framework (https://docs.nvidia.com/bionemo-framework/1.10/models/equidock.html).

18. TensorRT SDK | NVIDIA Developer (https://developer.nvidia.com/tensorrt).

19. NVIDIA cuEquivariance Documentation — cuEquivariance (https://docs.nvidia.com/cuda/cuequivariance/).

20. admin. Custom computing for the next generation of protein design – Institute for Protein Design. (2026). (https://www.ipd.uw.edu/2026/06/custom-computing-for-the-next-generation-of-protein-design/)

21. Palaniappan, L. P. Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. *Circulation* **153**, (2026).

22. Teo, K. K. & Rafiq, T. Cardiovascular Risk Factors and Prevention: A Perspective From Developing Countries. *Can J Cardiol* **37**, 733–743 (2021).

23. An overview of cardiovascular-kidney-metabolic syndrome. *Am J Manag Care* **30**, 181–188 (2024).

24. Harrold, J. M., Ramanathan, M. & Mager, D. E. Network-Based Approaches in Drug Discovery and Early Development. *Clin Pharmacol Ther* **94**, 651–658 (2013).

25. Erdmann, J., Kessler, T., Munoz Venegas, L. & Schunkert, H. A decade of genome-wide association studies for coronary artery disease: the challenges ahead. *Cardiovasc Res* https://doi.org/10.1093/cvr/cvy084. (2018) doi:10.1093/cvr/cvy084.

26. Hariharan, P. & Dupuis, J. Mapping gene and gene pathways associated with coronary artery disease: a CARDIoGRAM exome and multi-ancestry UK biobank analysis. *Sci Rep* **11, 16461**, (2021).

27. Cui, J. Recent advances in deep learning for protein-protein interaction: a review. *BioData Min* **18**, 43 (2025).

28. Alkhateeb, N. & Awad, M. Advances in protein-protein interaction prediction: a deep learning perspective. *Front Bioinforma* **5**, 1710937 (2026).

29. The UniProt Consortium *et al.* UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).

30. Hetmann, M. *et al.* Folding the human proteome using BioNeMo: A fused dataset of structural models for machine learning purposes. *Sci. Data* **11**, 591 (2024).

31. Ahdritz, G. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nat Methods* **21**, 1514–1524 (2024).

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

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

34. Tam, C. & Iwasaki, W. AlphaCutter: Efficient removal of non-globular regions from predicted protein structures. *PROTEOMICS* **23**, 2300176 (2023).

35. Barhoon, M. & Mahdiuni, H. Exploring Protein–Protein Docking Tools: Comprehensive Insights into Traditional and Deep-Learning Approaches. *J Chem Inf Model* **65**, 6446–6469 (2025).

36. Ganea, O.-E. Independent SE(3)-Equivariant Models for End-to-End Rigid Protein Docking. (2021) doi:10.48550/ARXIV.2111.07786.

37. NVIDIA H100 GPU Whitepaper (https://resources.nvidia.com/en-us-hopper-architecture/nvidia-h100-tensor-c).

38. Xue, L. C., Rodrigues, J. P., Kastritis, P. L., Bonvin, A. M. & Vangone, A. PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics* **32**, 3676–3678 (2016).

39. Liu, Z. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405–412 (2015).

40. Luck, K. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).

41. Oughtred, R. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* **30**, 187–200 (2021).

42. Szklarczyk, D. The STRING database in 2025: protein networks with directionality of regulation. *Nucleic Acids Res* **53**, 730–737 (2025).

43. Wang, R., Fang, X., Lu, Y. & Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *J Med Chem* **47**, 2977–2980 (2004).

44. Yoo, A. B., Jette, M. A. & Grondona, M. SLURM: Simple Linux Utility for Resource Management. in Job Scheduling Strategies for Parallel Processing. vol. 2862 44–60 (2003).

45. Ruiz-Carmona, S. rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput Biol* **10**, 1003571 (2014).

46. Comajuncosa-Creus, A., Jorba, G., Barril, X. & Aloy, P. Comprehensive detection and characterization of human druggable pockets through binding site descriptors. *Nat Commun* **15**, 7917 (2024).

47. Caldera, M., Buphamalai, P., Müller, F. & Menche, J. Interactome-based approaches to human disease. *Curr Opin Syst Biol* **3**, 88–94 (2017).

48. Shin, C. J., Wong, S., Davis, M. J. & Ragan, M. A. Protein-protein interaction as a predictor of subcellular location. *BMC Syst Biol* **3**, 28 (2009).

49. Wang, K. Transcriptome analysis of muscle atrophy in Leizhou black goats: identification of key genes and insights into limb-girdle muscular dystrophy. *BMC Genomics* **26**, 80 (2025).

50. McCarthy, M. I. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9**, 356–369 (2008).

List of Figures:

Figure 1: **GPU-accelerated proteome-wide PPI prediction framework.** Overview of an AI-driven docking pipeline screening 126 million human protein pairs. Utilizing EquiDock across 460 H100 GPUs achieves a 300–2000x speedup over traditional tools, identifying 56.1 million novel interactions

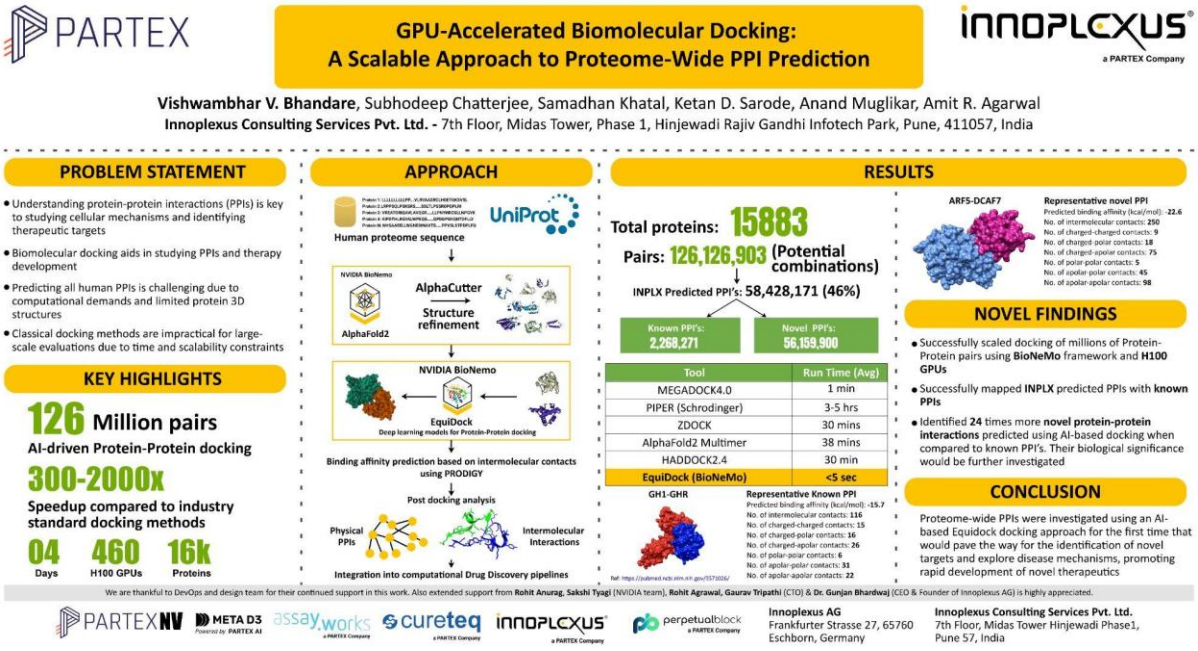


Figure 2: Integrated computational and genetic prioritization pipeline for coronary artery disease (CAD) interactome discovery. The schematic workflow illustrates the three-tiered end-to-end framework developed to systematically investigate the mechanistic organization of CAD at a systems level by bridging sequence-based interactomics with large-scale population genetics.

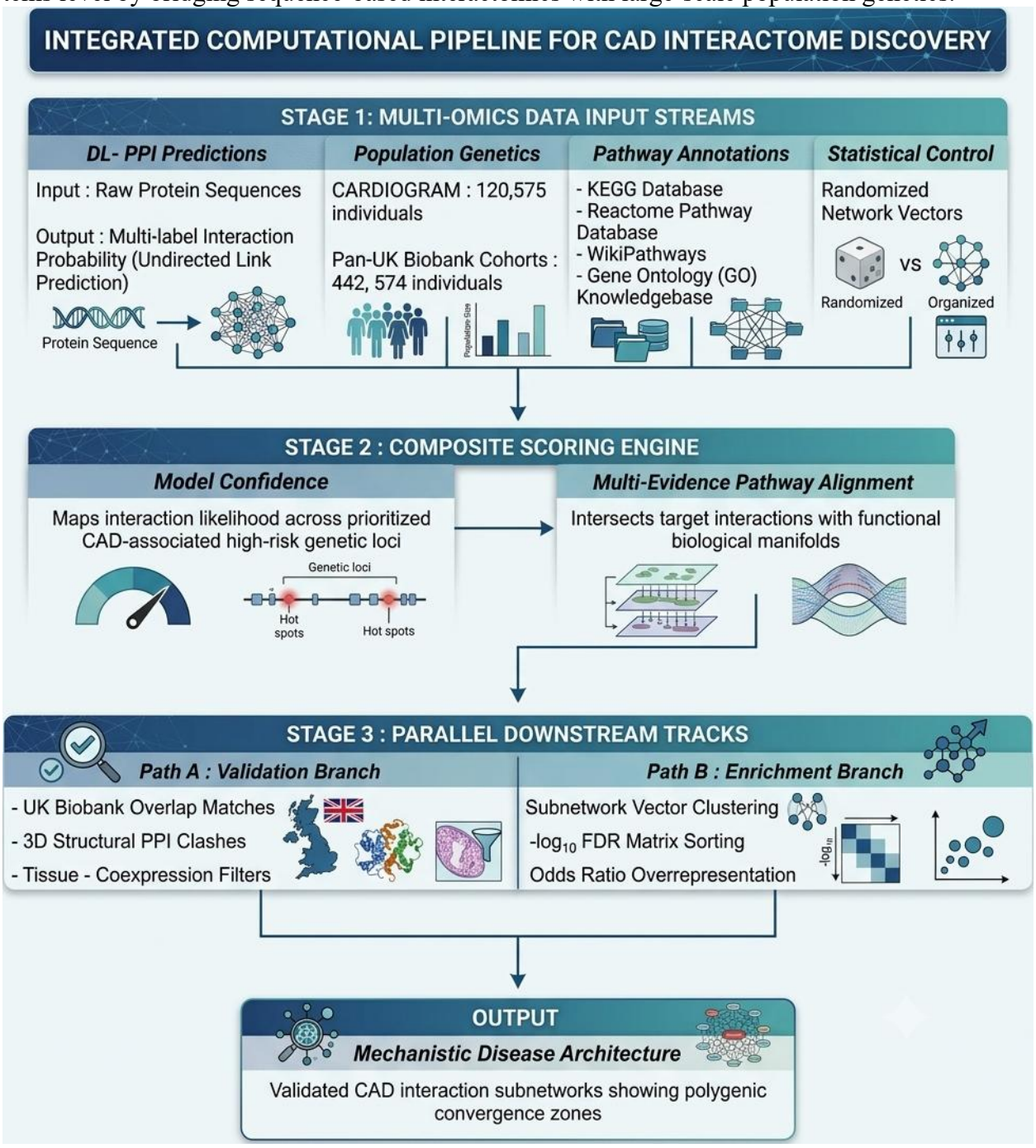


Figure 3: **Architectural workflow of the automated POcketVec -driven structural refinement pipeline.**
Raw PDB coordinate preparation proceeds through four sequential stages: ensemble reduction to Model 0 using Bio.PDB, functional polymer chain isolation via ChainSelect, heteroatom matrix filtering (waters, ions, and crystallization solvents), and final conversion to Tripos MOL2 format via OpenBabel with Gasteiger charge assignment and atom typing.

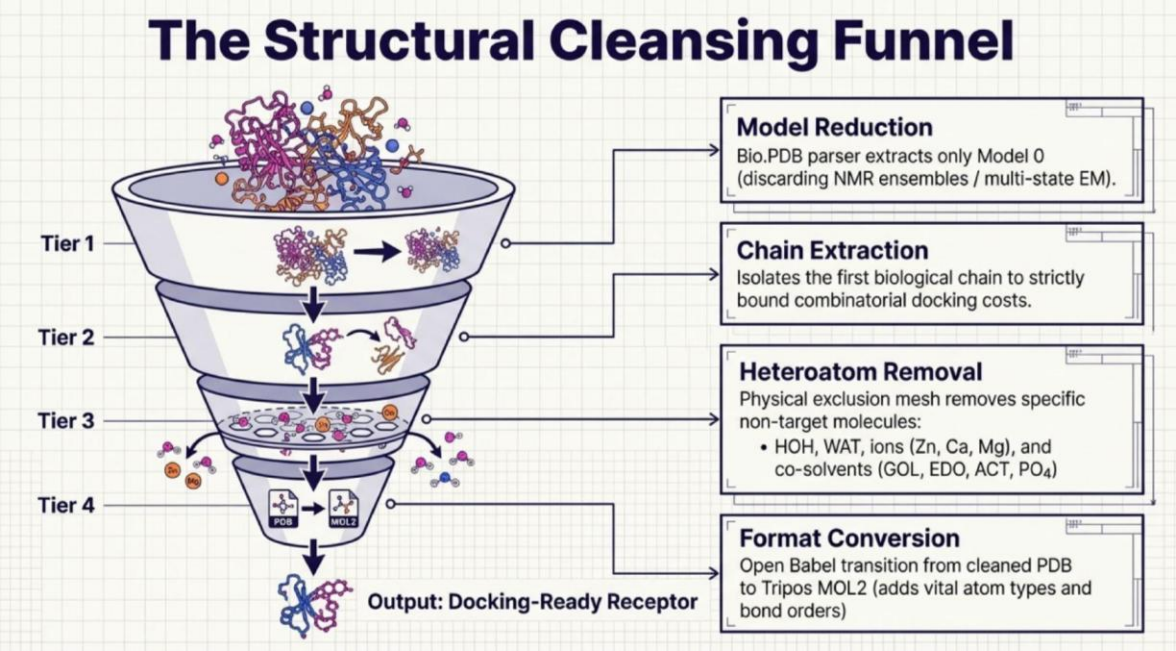


Figure 4: Composite-score interaction matrix revealing polygenic convergence within mitochondrial respiratory complexes.
The symmetric heatmap displays the pairwise interaction topology for the top 50 high-confidence predicted protein-protein interaction (PPI) pairs across 23 prioritized coronary artery disease (CAD)-associated genes. Both vertical (Y-axis) and horizontal (X-axis) axes are ordered identically to profile structural and functional cross-talk across the candidate proteomic space. Individual cell color intensity corresponds to the programmatic Composite Score (CS), mapped along a continuous gradient from dark purple/black (CS = 0.0; no functional interaction interface detected) to bright teal and soft mint green (CS < 0.5; maximum structural and multi-evidence biological alignment).This network architecture illustrates a molecular mechanism for polygenic convergence in CAD, demonstrating how distributed, sub-threshold genetic variations cluster within physical, interconnected macromolecular complexes to drive tissue-specific metabolic dysfunction.

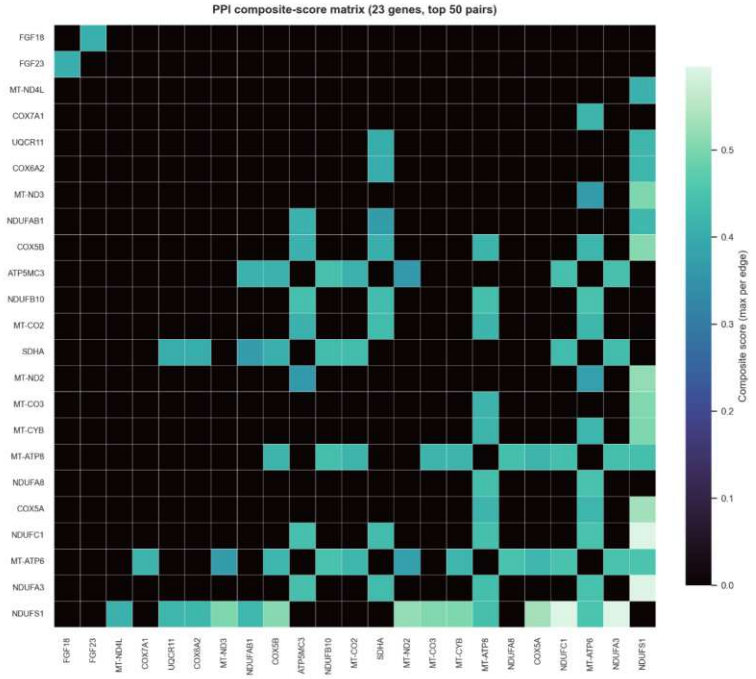


Figure 5: Multi-evidence pathway overlaps heatmap for the top 50 predicted PPI pairs. The heatmap profiles the functional annotation depth and cross-database support for the highest-confidence predicted protein-protein interaction (PPI) pairs.

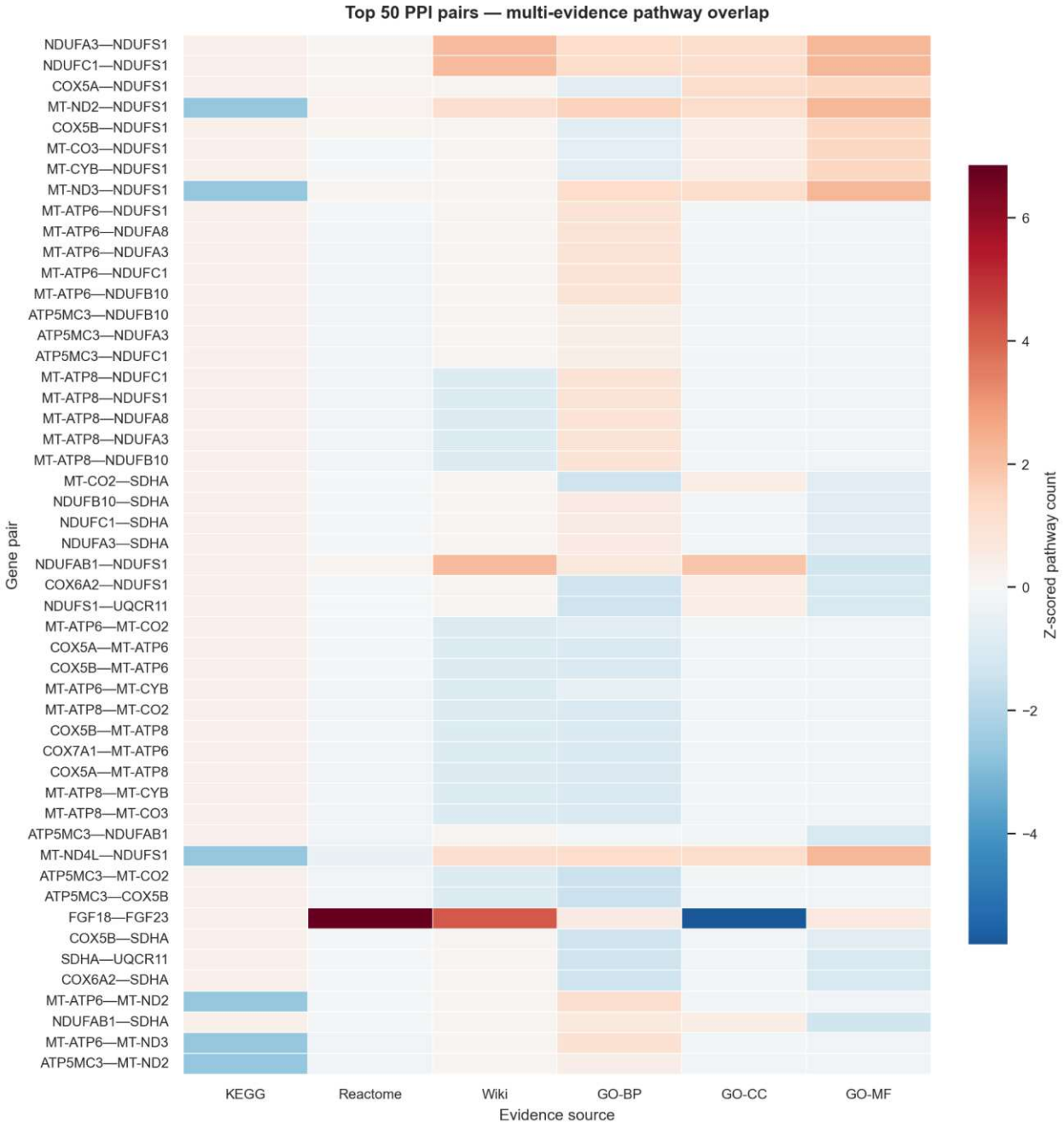


Figure 6: Functional overrepresentation of Gene Ontology (GO) Biological Processes within prioritized interactome modules.
The bubble plot details the statistical enrichment of GO Biological Processes within the top 10% ranked protein-protein interaction (PPI) pairs compared against the remaining proteomic background.

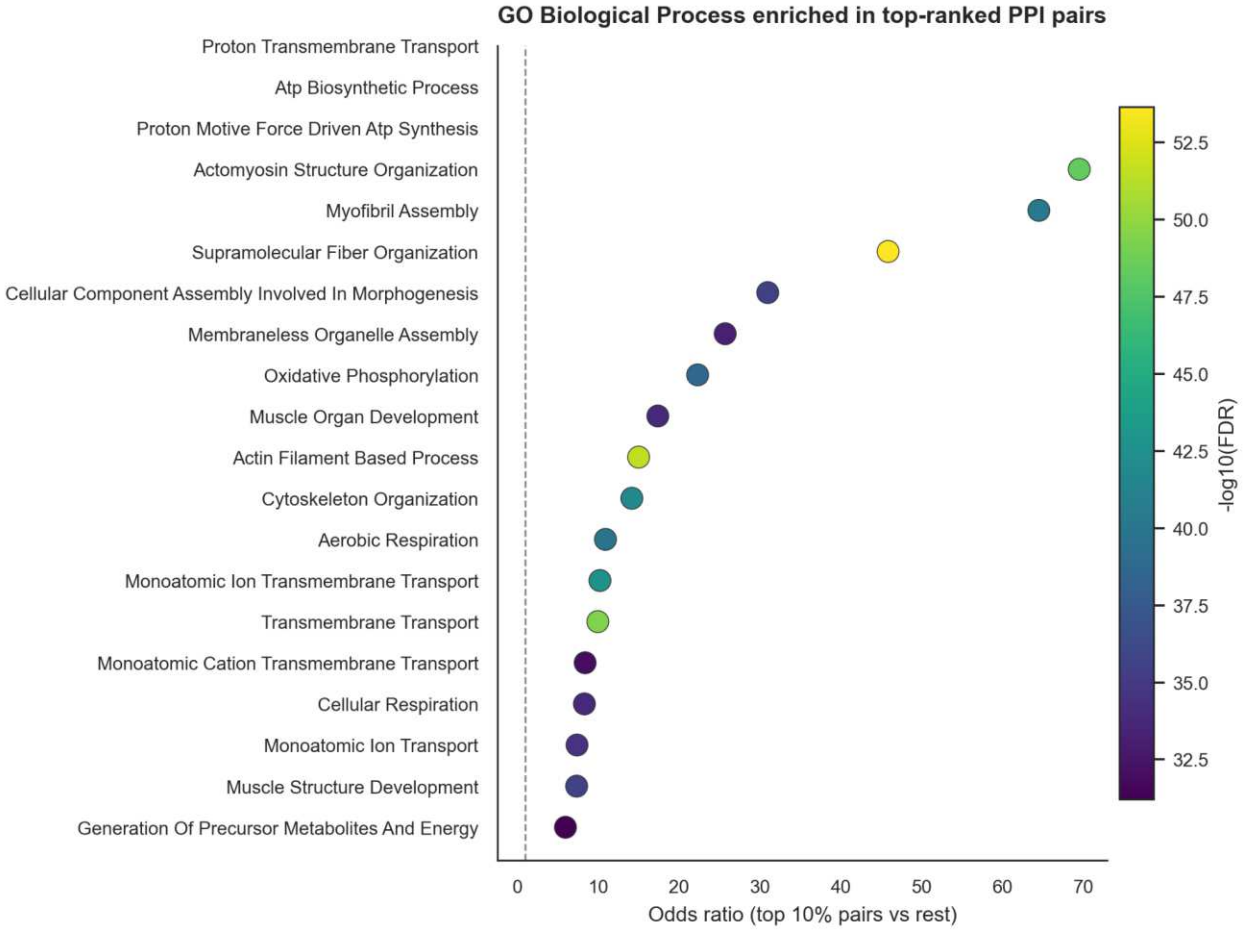


Figure 7: **KEGG pathway overrepresentation profiling for top-ranked CAD-associated PPI pairs.**
The bubble plot details the statistical enrichment of canonical KEGG pathways within the top 10% ranked protein-protein interaction (PPI) pairs relative to the remaining proteomic background space.

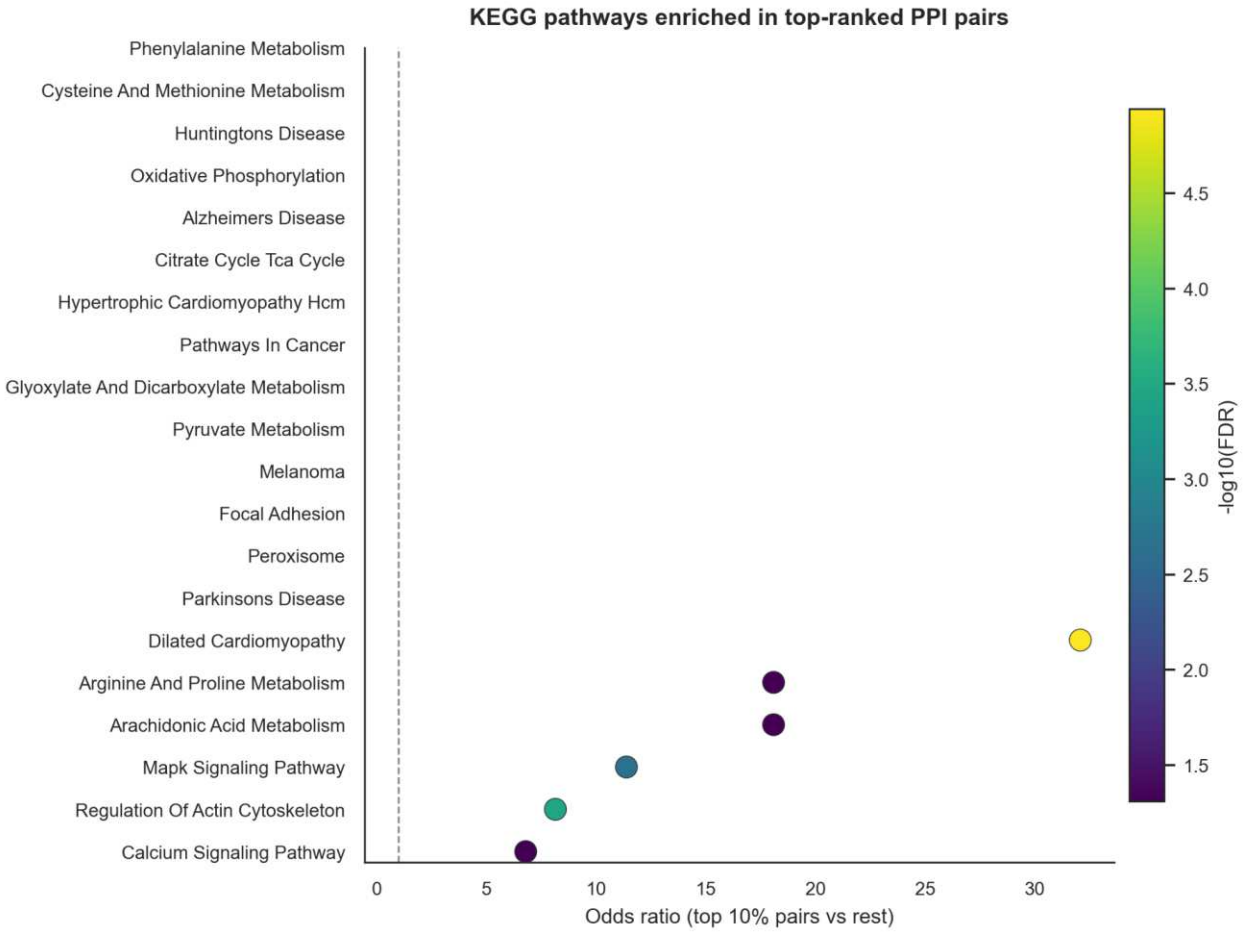


Figure 8: Reactome pathway overrepresentation profiling for top-ranked CAD interactome pairs.
The bubble plot displays the statistical overrepresentation of high-level and granular Reactome pathways within the top 10% ranked protein-protein interaction (PPI) pairs relative to the remaining background proteome.

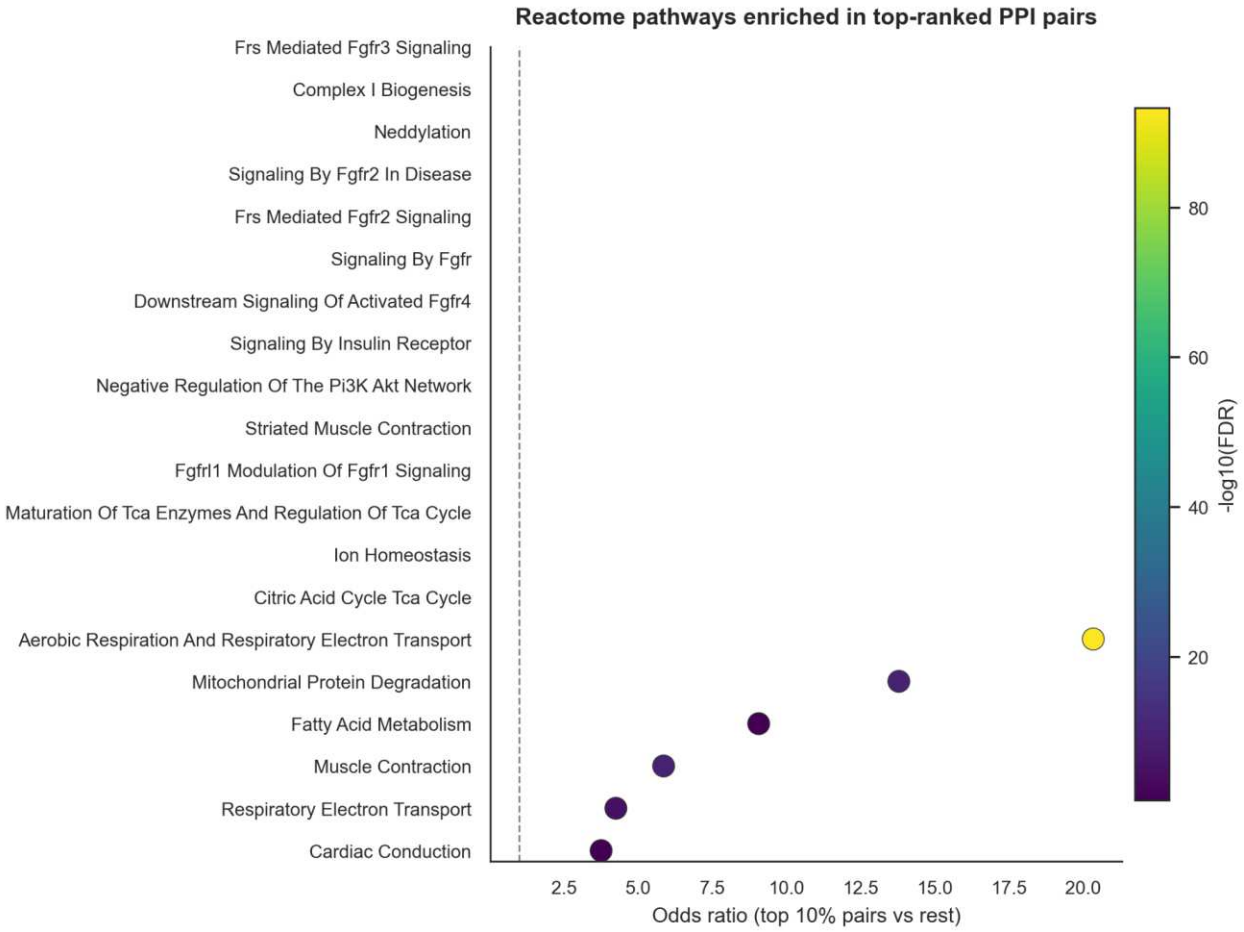


Table 1: Number of protein, PPI pairs and systemic filtering criteria used to identify novel PPI's.

Filtration Stage	Evaluation Unit	Total no. of proteins	Combinatorial Space (Unique PPI Pairs)
Theoretical Proteomic Input	Unfiltered UniProt IDs	20,505	115,000,000
Cleaned Functional Curation	Removed structural junk/fragments	17,112	113,000,000
Reviewed Reference Cohort	High-confidence Human Entries	16,164	115,000,000
Subcellular Co-Localisation	Stripped ~1,700 missing addresses	16,164	27,516,110
Unlocalised Structural Matrix	AA < 1000, Loop < 7 (No localisation)	6,005	15,571,302
Co-Localization filter	Dual Structural & Spatial Restricted	6,005	3,314,683

Table 2: Mapping of predicted PPI's with known PPI's databases.

Database	Total entries	Human entries	Mapped pairs
BioGRID	321,476	302,747	86,131
HuRI	1,280,687	54,864	54,506
PDBBind	415	415	415
STRING	1,182,858	474,605	467,412
Unified Master Benchmark	-	-	529,492 (Filtered), 91,883 (Structural Sub-set)

Table 3: Mapping against filtered, annotated PPI's database subsets

Databases	Reference Pairs	Total mapped pairs (from 15.5M)	Mapping %	Mapped pairs (from 3.3M Matrix)	Identified novel PPI's (in % from 3.3M pairs)
BioGRID	12,823	10,849	84.61%	7,879	61.44%
HuRI	10,373	8,641	83.30%	3,356	32.35%
PDBBind	64	64	100.00%	45	70.31%
STRING	81,335	70,651	86.87%	38,408	47.22%

Table 4: Identified high confidence novel PPI's using subcellular localization filtering.

PPI corpus	Total Pairs	Mapped Pairs	Mapping %	Unique Novel Interaction
Structural Matrix (Full 529K database)	15,571,302	79,335	14.98%	15,491,967
Co-Localized Structural Matrix (Full 529K database)	3,314,683	42,051	7.94%	3,272,632
Structural Matrix (Filtered 91K Subset)	15,571,302	79,175	86.17%	3,235,508
Co-Localised Structural Matrix (Filtered 91K Subset)	3,314,683	42,051	45.76%	3,272,632

Table 5: Cardiovascular (CVD) space associated novel PPI's.

CVD associated PPI filtering	Dataset used	Number of PPI's	Novel PPI %
Initial Cardiovascular Space	Unfiltered Input Pairs	85,000	Baseline
Structural & Loop constraints	AA < 1000 residues, Loops < 7 AA residues	11.026	12.97% Structural Yield
First-Pass Database Overlap	BioGRID, HuRI, PDBBind, STRING	10,814 (Novel), 212 (Known)	98.085 Preliminary Novelty
Second-pass Overlap	Secondary Clean Vs. STRING/BioGRID	10,641 (Novel), 173 (Known)	96.50% Net Structural Novelty
Energy Constraints Mapping	Filtered Vs. 115M Energy Database	9,624	90.44% Thermally Stable
High-confidence Candidate Selection	Top 10% Highest-Scoring Energies	962	8.72% Final Target Yield

Table 6: Functional Curation of predicted PPI's based on Tissue-Specific Interactome

Filtering rule	Filtering criteria	Total Pairs	Novel PPI's %
Tissue-Structural Intersection	Co-expressed Structural Matrix	1,089,656	100.00%
First-pass Database Check	Standard Master Overlap	1,085,419 (Novel)	99.61%
Second-pass Reference Check	High-confidence Stringent Databases	1,076,657 (Novel)	98.80% Net Expansion Yield
Functional Pathway Filter	Ontology Network Mapping	445,676	41.39% of Novel Dataset
Text-Mining Validation FIlter	PubMed Text Matrix Intersect	436,841	40.57% Highly Prioritised Target Pool

Table 7: A list of top 20 prioritized cardiovascular associated genes / targets across VEGAS2 and MAGMA pipelines

ADAMTS7	APOE	LPA	SH2B3	LPL
PCSK9	HNF1A	FAM177B	CEKSR2	MYBPHL
ANGPTL4	DUSP13	EHBP1L1	KIAA1462	ARHGEF26
LAMB2	FN1	UNC5C	TNS1	UHRF1BP1

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFilenaturecardio.docx