bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

A Framework for Autonomous AI-Driven Drug Discovery

Douglas W. Selinger*[†], Timothy R. Wall[†], Eleni Stylianou[†], Ehab M. Khalil[†], Jedidiah Gaetz[†], Oren Levy[‡]

[†]Plex Research Inc., Cambridge, MA, USA; [‡]Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

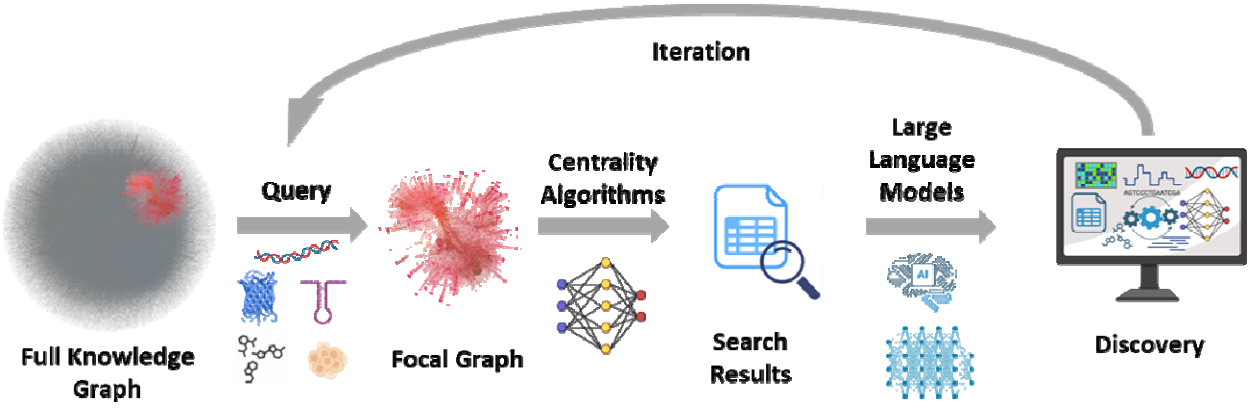
***Corresponding author (dselinger@plexresearch.com), www.plexresearch.com**

Abstract

The exponential increase in biomedical data offers unprecedented opportunities for drug discovery, yet overwhelms traditional data analysis methods, limiting the pace of new drug development. Here we introduce a framework for autonomous artificial intelligence (AI)-driven drug discovery that integrates knowledge graphs with large language models (LLMs). It is capable of planning and carrying out automated drug discovery programs while providing details of its research strategy, progress, and supporting data points, enabling a thorough assessment of its methods and findings. At the heart of this framework lies the “focal graph” - a novel construct that harnesses centrality algorithms to distill vast, noisy datasets into concise, transparent, data-driven hypotheses. By enabling high-throughput search and automated result interpretation, such a framework could be used to execute massive numbers of searches, identify patterns across complex, diverse datasets, and prioritize actionable hypotheses at a scale and speed unachievable by human researchers alone. We demonstrate that even small-scale applications of this approach can yield novel, transparent insights relevant to multiple stages of the drug discovery process and present a prototype system capable of autonomously planning and executing a multi-step target discovery workflow. The focal graph framework described here, and the automation it enables, represents a promising path forward: towards a deeper understanding of the mechanisms underlying disease and a true acceleration in the development of novel therapeutics.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Graphical Abstract



Introduction

In recent decades, biomedical research has entered an era marked by unprecedented data generation. Advances in multi-omics, chemical biology, and high-throughput screening have led to a deluge of information, creating a situation where, paradoxically, the abundance of data actually hinders progress. This trend is exemplified by Eroom's Law, which observes that drug discovery has become slower and more expensive over time, despite technological progress.[1,2] At the same time, the vast stores of untapped biomedical data offer a potential solution to this trend,[3] yet the sheer volume, noise, and complexity of these datasets have rendered them largely inaccessible through traditional research approaches.[4,5] There is tremendous interest in the potential of AI-driven solutions to overcome these barriers, however many challenges remain.[6,7]

Here we propose a framework for autonomous AI-driven drug discovery, capable of navigating and integrating large-scale, complex datasets to uncover novel scientific insights. At the core of this system is the "focal graph," a novel type of knowledge graph optimized for scalability and interpretability.[8] Unlike traditional knowledge graphs, which grow unwieldy as they expand, focal graphs leverage centrality algorithms, such as PageRank, to focus on highly connected subregions within larger networks.[9] This targeted approach enables the system to identify connections between key nodes - such as compounds, genes, or biological pathways - relevant to specific research queries, facilitating efficient data extraction even from vast, noisy datasets.

Building on this foundation, we will show that such a system can employ large language models (LLMs) to autonomously plan, execute, and interpret focal graph-based searches. LLMs can function as both search agents and interpreters, capable of generating hypotheses, refining search strategies, and summarizing findings across thousands of queries. We propose that this integrated approach allows continuous, iterative discovery, where new insights inform subsequent searches, with the results can synthesized into actionable conclusions. The

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

autonomous capability of the system has the potential to scale effortlessly, transforming the biomedical data deluge into a torrent of structured, evidence-backed insights.

Here, we demonstrate the application of this focal graph-based, AI-driven system to multiple stages of the drug discovery process. By examining cases where focal graph analyses reveal novel insights, we show that even small-scale applications yield data-supported hypotheses. Moreover, we illustrate how scaling this approach across larger datasets amplifies its power, making it feasible to conduct research programs that would be impractical for human scientists alone. As data-driven methodologies become essential in modern drug discovery, our approach offers a promising path toward autonomous, AI-driven drug discovery—one that combines AI scalability with the transparency and rigor required for scientific discovery.

Focal Graphs

Knowledge graphs are structured representations of real-world entities, with the entities represented as nodes, e.g., circles, and their relationships represented as edges, e.g., lines or arrows. While they have become increasingly prevalent in drug discovery and biomedical research,[10–12] their utility is often limited by computational complexity and interpretability challenges when applied to complex diseases and drug mechanisms.[13] As we will show, focal graphs address these limitations in multiple ways.

Focal graphs are subgraphs of knowledge graphs, constructed such that the properties of their seed members, or “queries”, can be inferred from their graph properties. Many kinds of focal graphs can be constructed to answer many different kinds of questions. For example, focal graphs constructed based on the known targets of structurally related compounds can be used to infer the putative target(s) of a seed compound.[8,14] They can be adapted to many other applications by modifying their construction and analysis, for example by varying the seed type, the knowledge graph from which they are derived, or by optimizing them to infer a particular property or set of properties. This makes them an extremely flexible tool for extracting many different kinds of insights.

Starting from a query of interest - whether a compound, gene list, or disease signature - a focused subgraph is constructed that captures the most relevant relationships within the larger knowledge network (Figure 1A). This focused extraction enables detailed analysis of specific research questions while maintaining computational tractability, even as the underlying knowledge graph grows arbitrarily large. Centrality algorithms are mathematical methods used to determine the relative importance or influence of nodes within a graph based on their connectivity and position. Their application to focal graphs yields concise, rank-ordered lists of the most highly connected graph members, which can be used in turn to infer the potential properties of the original queries.

The construction of a focal graph follows a systematic workflow (Figure 1A). First, entities similar to the query are identified, creating an initial layer of relationships. These similar entities are then connected to their associated properties - such as compounds, targets, pathways, or disease associations - forming a second layer of connectivity. The resulting multi-layered

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

network can then be analyzed using centrality algorithms, such as PageRank, which have proven effective in identifying essential nodes in biological networks.[15] Centrality algorithms identify results with the broadest levels of support, making them highly robust to noise. They also focus attention on the most strongly supported results, making them highly concise and scalable. The most experimentally supported results are ranked first, regardless of how much data is searched or how many additional results are returned. In fact, the larger and more diverse the global knowledge graph becomes, the better the focal graph-derived results should be, without sacrificing robustness, concision, transparency, or simplicity.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

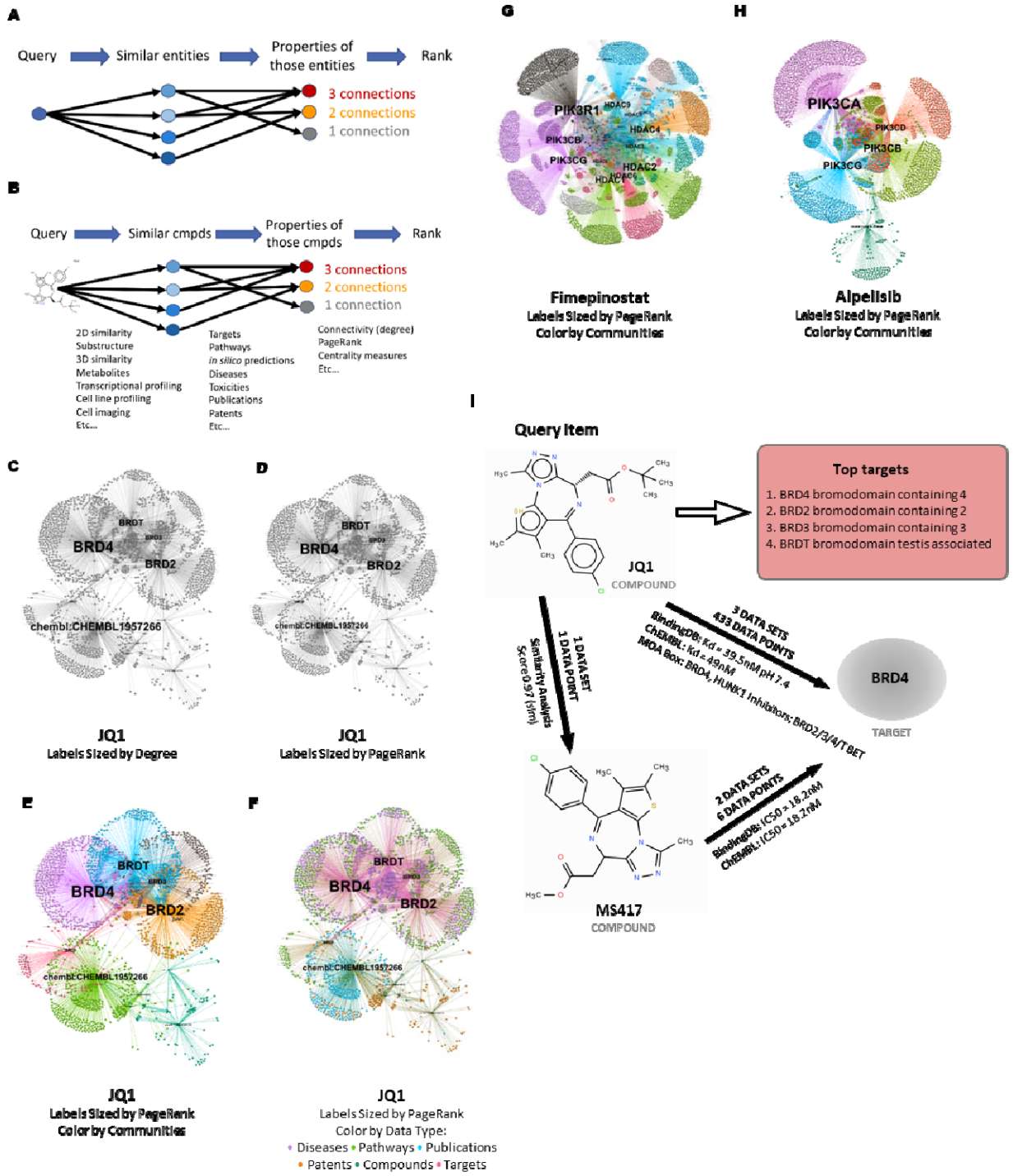


Figure 1. Overview of focal graphs *A) General focal graph-building workflow showing a query node (leftmost blue circle) linked by edges (represented by arrows) to nodes representing similar entities (middle circles in shades of blue), then linked to nodes representing their properties (rightmost circles), which are then ranked by a graph centrality algorithm, such as degree of connectivity. B) Focal graph building workflow for compound target ID. The query example here, JQ1, is a compound structure from which similar compounds can be found and additional information on its targets, pathways, disease indications and other properties can be inferred. C,D) A focal graph generated for JQ1, a BET family inhibitor with known targets BRD4, BRD3, BRD2, & BRDT, with labels sized by*

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

degree (C) or by PageRank (D). E,F) Focal graphs for JQ1 with labels sized by PageRank and colored by graph community (E) or by data type (F). G) Focal graph of Fimepinostat, a pan-PI3K and pan-HDAC inhibitor, with labels sized by PageRank and colored by graph community. H) Focal graph of Alpelosib, an alpha specific PI3K inhibitor, with labels sized by PageRank and colored by graph community. I) Selected focal graph evidence paths that connect JQ1 to BRD4, one of its known targets through binding data and chemical similarity. Focal graph images in C-H were generated by Gephi (https://gephi.org).

Focal graphs are exceptionally robust to noise. Datasets with high levels of random noise will fail to agree with other data sources, pushing their results lower in the rankings. Datasets with systemic noise will only agree with other data with the same bias, but not with other, independent approaches, pushing their results lower in the rankings while also potentially making the source of the systematic noise easier to identify.

In contrast to many machine learning (ML) based methods, the data supporting focal graph-derived results can be easily identified and traced back to their original sources, allowing in-depth evaluation by subject matter experts. This high level of transparency is critical for scientific and regulatory applications where specific details of the supporting data is crucial for decision-making. This combination of properties makes focal graphs ideally suited for dealing with the huge explosion of complex, large scale datasets that are ubiquitous in modern biomedical research.

The power of focal graphs becomes evident when examining specific examples. A focal graph generated using centrality algorithms for the BET bromodomain inhibitor JQ1 (Figure 1B), a well-characterized chemical probe,[16] reveals distinct communities of interconnected nodes sized by degree (Figure 1C), by PageRank (1D), or colored by graph community (1E) or data type (1F), each representing different aspects of the compound's biological activity. These graph communities, or neighborhoods, share similar connections and potentially also biological functions, and can be excised and studied independently. Community structure can be driven by many factors including common chemical structure, biological processes, data provenance, bioactivity, or some combination of these.

When colored by data type (Diseases, Pathways, Publications, Patents, Compounds and Targets, Figure 1F), we can observe how different types of evidence contribute to an understanding of the compound's mechanism of action. Generating a focal graph using the structure of JQ1 as a query identifies connections to its known molecular targets, including BRD4, BRD2, BRD3 and BRDT (Figure 1I). This illustrates how focal graphs can integrate both direct and indirect supporting data, e.g., chemical similarity algorithms and data from diverse sources such as ChEMBL and PubChem.[17,18] Focal graphs constructed for Fimepinostat, a pan-PI3K and pan-HDAC inhibitor (Figure 1G),[19] and Alpelisib, an alpha specific PI3K inhibitor (Figure 1H),[20] reflect each compound’s known, distinct polypharmacology.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Synergy with Machine Learning Approaches

While machine learning approaches have demonstrated impressive predictive capabilities in drug discovery,[21] they often face challenges with data heterogeneity, lack of transparency, result interpretability, and the need for large training sets.[22] Neural networks in particular, including graph neural networks and deep learning models, typically operate as "black boxes," making it difficult to trace the evidence supporting their predictions.[23] In contrast, focal graphs preserve the provenance of their insights, allowing researchers to examine the exact experimental data and relationships that support each conclusion.

Many of the edges of the focal graphs in Figures 1-3 were created at run time by chemical similarity algorithms. Edges can in principle be generated by any kind of algorithm, including ML-based predictions,[24–26] allowing focal graphs to simultaneously integrate diverse data sets and *also* a variety of algorithms, identifying for example, where experimental data and predictive algorithms most converge. This can help optimize algorithms by identifying where they agree, or fail to agree, either with other algorithms or with experimental data. Agreement with experimental data may provide a broader view of the biological or chemical context that drove the algorithmic prediction, making it more interpretable and potentially yielding insights that may help refine the predictive approach.

The robustness of focal graphs stems from their ability to identify consensus across multiple data sources and experimental approaches. Unlike supervised learning methods that may learn spurious correlations from training data,[27] focal graph centrality algorithms naturally highlight instances where multiple independent lines of evidence converge on the same conclusion. This approach is particularly valuable in drug discovery, where experimental noise and systematic biases are common challenges.[28]

Focal graphs also offer distinct advantages in handling sparse or novel data. While machine learning models struggle with predictions outside their training distribution, focal graphs can surface meaningful relationships for novel entities by connecting them to a small number of biologically or chemically related examples. In fact, even a single strong connection can be enough for a focal graph to generate a compelling hypothesis, whereas a ML-based model typically takes a large number of examples to train and test. This capability is especially relevant for emerging therapeutic modalities or newly discovered disease mechanisms where training data may be limited.

The relationship between focal graphs and machine learning approaches is more complementary than competitive. Machine learning can be incorporated into focal graph construction strategies by the inclusion of computationally created edges. Focal graphs can provide interpretable context for machine learning predictions by connecting them to relevant experimental data as well as predictions from other computational approaches.[29] Furthermore, focal graphs can themselves be used to generate training sets for machine learning models, for example by providing rank-ordered consensus examples for a variety of query types, such as a set of compound structures most associated with inhibition of a particular target or a set of genes most associated with a toxicity. Thus, the integration of focal graphs with ML-based

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

prediction, particularly when combined with large language models, can lead to natural synergies that produce novel findings while maintaining scientific rigor and transparency.

Focal Graph - Large Language Model (FG-LLM) Integration

Large language models (LLMs) such as ChatGPT or Claude have a remarkable ability to converse in natural language, incorporating an impressive breadth of information gleaned from massive text-based training sets. Expanding their capabilities beyond text-based data is an area of active investigation.[30,31] Combining LLMs with an internet search engine has proven to be a highly successful and easily implemented strategy.[32] However, internet searches do not currently include the experimental results from large scale biomedical datasets, such as chemical structure/bioactivity data or multi-omics datasets, and it remains unclear the extent to which LLMs can interpret chemical structures in a way that enables chemical similarity or substructure searching.[33,34] A focal graph search is an effective way to connect these data to LLMs, as demonstrated in the use cases described below, potentially reducing the frequency of hallucinations and dramatically enhancing the power of LLMs to advance drug discovery.

Integration of focal graphs with large language models allows the automatic generation, execution, and interpretation of searches, allowing them to be included as part of large scale "intelligent" workflows. These automated searches can be run serially for many iterations, potentially allowing execution of an extended computational research program that uses massive amounts of existing data - all while maintaining full transparency with respect to the research strategy and the data that supports the hypotheses uncovered. Key findings of large scale workflow runs can be summarized for easy evaluation and provide the basis for larger scale target prioritization efforts. We will return to this topic and cover it in more detail later.

Diverse Applications of Focal Graphs to Drug Discovery

Focal graphs can capture complex relationships between biological entities or nodes, making them applicable to the exceptionally wide range of datasets characteristic of modern biomedical research. This allows them to find connections between almost any data type and makes them applicable to a wide range of experimental approaches, algorithms, and scientific questions.[8,14,35]

Here we present several case studies that demonstrate the versatility and power of focal graphs. We show how large language models can be employed to systematically interpret their results and suggest new avenues for investigation. The examples presented here, while focused on specific use cases, illustrate general principles that can be applied to a broad range of drug discovery challenges.

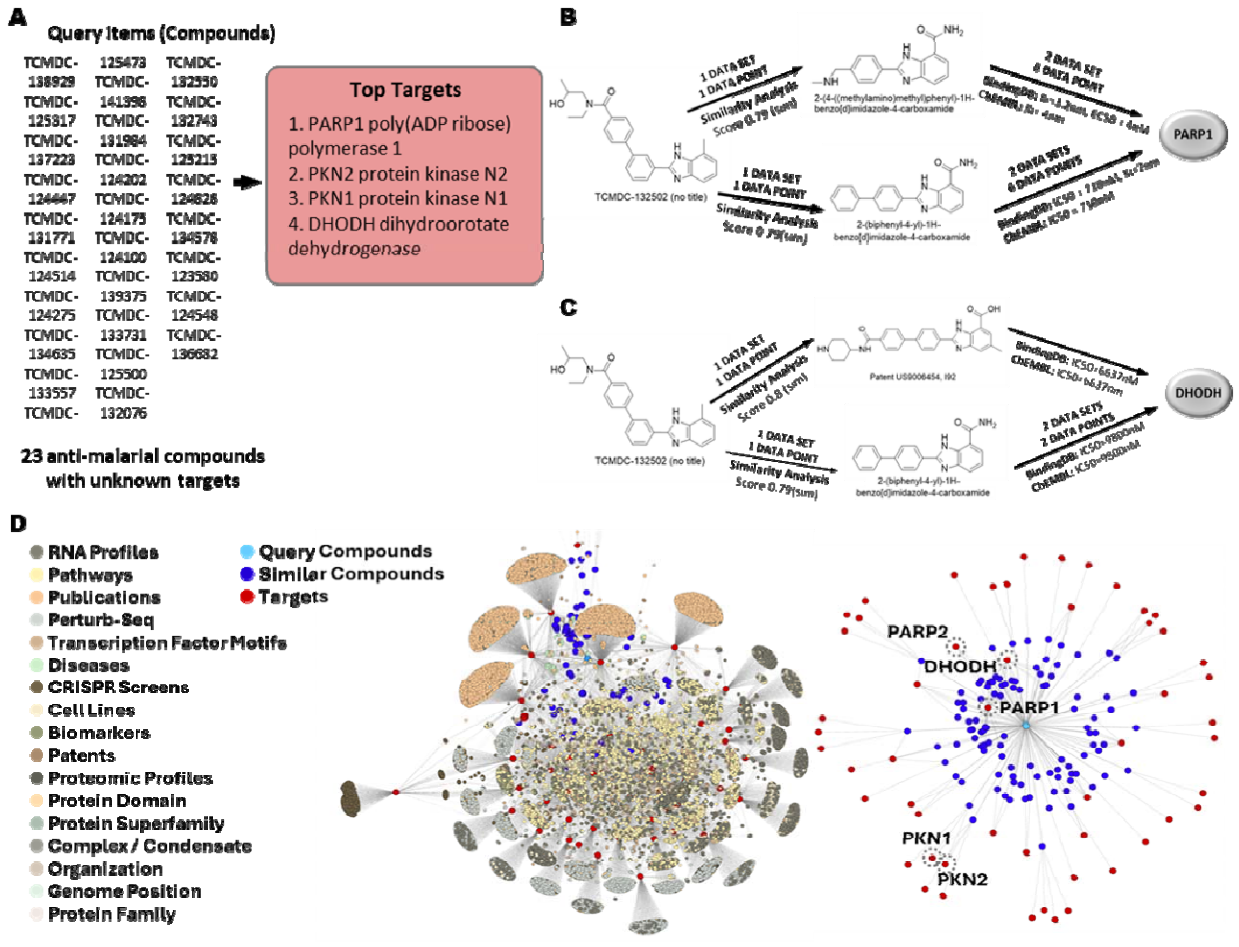
bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Chemistry-Driven Target Discovery

Phenotypic screening is widely used as an approach to discover highly relevant modulators of biological processes for drug development." However, identifying the molecular targets of the screening hits and the mechanisms involved, remains a difficult challenge.[36–39]

We used a focal graph to analyze 23 compounds from a chemical series (series 46) identified by Calderon et al. as having antimalarial activity but whose target(s) are unknown (Figure 2).[40] A focal graph search of the compound series revealed two highly ranked potential targets: poly(ADP-ribose) polymerase 1 (PARP1) and dihydroorotate dehydrogenase (DHODH), each supported by multiple lines of evidence (Figure 2A-D). While the focal graph search identified more high potency series 46-analogs connected to PARP1 than to DHODH, the LLM analysis found that DHODH was more strongly connected to malaria, with data coming from multiple sources. These included the connection to the DHODH inhibitor DSM265, an investigational antimalarial treatment that progressed to phase II clinical trials.[41,42] This broader, cross-data type perspective led the LLM to conclude that the antimalarial activity of the series was best explained by DHODH inhibition (Figure 2E). Notably, the focal graph identified a publication that described a structurally related series as having dual PARP1 and DHODH activity,[43] suggesting that series 46 may in fact inhibit both targets.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.



E

Focal Graph-Large Language Model (FG-LLM)

This search was run with a list of compound structures that have antimalarial activity but whose target is unknown. Based on a careful review of the results, what do you think the target might be that is driving the antimalarial activity? Please include specific supporting evidence.

After carefully reviewing the search results for the list of compound structures with antimalarial activity, a strong candidate for the target driving the antimalarial activity appears to be **dihydroorotate dehydrogenase (DHODH)**. There are several pieces of evidence supporting this hypothesis:

1. **High ranking target:** DHODH is ranked third in the Target category, indicating strong relevance to the query compounds.
2. **Antimalarial connection:** DHODH is known to be a target for antimalarial drugs. This is supported by the publication: *"Potent Antimalarials with Dihydroorotate Dehydrogenase Inhibitory Activity: Structure-Guided Computational Optimization of a Pyrrole-Based Dihydroorotate Dehydrogenase Inhibitor Series"*, which directly links DHODH inhibition to antimalarial activity.
3. **Structural evidence:** The search results include several patents related to DHODH inhibitors, suggesting structural similarities between the query compounds and known DHODH inhibitors:
   - *"Dihydroorotate dehydrogenase inhibitors"*
   - *"2,4,5-trisubstituted 1,2,4-triazolones: useful as inhibitors of DHODH"*
   - *"Dihydroorotate dehydrogenase inhibitors"*
4. **Benzimidazole connection:** Many of the query compounds contain benzimidazole structures, and there's evidence linking benzimidazoles to DHODH inhibition. The patent *"Bis-benzimidazole compounds and methods of using the same"* suggests that benzimidazole-containing compounds may target DHODH.
5. **Research focus:** The publication *"Design and synthesis of potent substrate-based inhibitors of the Trypanosoma cruzi dihydroorotate dehydrogenase"* indicates ongoing research into DHODH inhibitors for parasitic diseases, which is relevant to malaria.
6. **Biological pathway relevance**: DHODH is involved in pyrimidine biosynthesis, which is crucial for rapidly dividing cells such as those of malarial parasites. Inhibiting this enzyme would disrupt the parasite's ability to replicate its DNA and proliferate.

While DHODH appears to be the most likely target driving the activity, other targets such as PARP1 (ranked first) and PKN2 (ranked second), may also play roles in the compounds' activity. However, the specific evidence linking these targets to antimalarial activity is less apparent in the provided search results.

To further confirm this hypothesis, additional experiments would be needed, such as enzymatic assays with recombinant DHODH and the query compounds, as well as growth inhibition studies in the presence of pyrimidine supplementation, which would rescue DHODH inhibition if it is indeed the primary target.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Figure 2. Focal graph analysis of a 23-member compound series, active in an antimalarial phenotypic screen. *A) The top ranked targets from a focal graph search of 23 compounds from an antimalarial phenotypic screen are shown. B,C) Selected focal graph evidence paths connecting a member of the compound series to PARP1 (B) or to DHODH (C) through chemical similarity and binding data. D) The focal graph of the 23 antimalarial compounds showing connections to many diverse data types. A subgraph of the focal graph is shown on the right comprising the connections between only the query compounds, similar compounds, and targets. The top ranked targets are labeled. E) Large Language Model (LLM) analysis of the focal graph search which highlights DHODH inhibition as the best-supported antimalarial target hypothesis.[44,45] Focal graph images in D were generated by Gephi (https://gephi.org). Focal graphs were used in combination with Claude 3.5 Sonnet, a foundational LLM from Anthropic.*

This example illustrates several key advantages of the focal graph approach. The strength of the structural similarity across the full series can be analyzed in detail and its consistency (or lack thereof) with previously established structure activity relationships can be assessed. Also, because the focal graph integrates extremely diverse data types, many aspects of the target hypothesis can be evaluated at once. This includes considering not only the molecular pharmacology but also the biology of the target, as described in the scientific literature and/or inferred directly from large scale data sets present in the focal graph. Ultimately, the transparency of the results allows expert evaluation of the full breadth of the underlying data and enables a nuanced analysis tailored to the specific goals and detailed context of the overall project.

Morphological Profiling

Cell painting, a high-dimensional imaging assay that captures cellular morphological states, offers an unbiased approach to characterization of a compound's mechanisms of action.[46–48] However, translating these rich phenotypic profiles into specific molecular hypotheses remains challenging, though significant progress with this methodology is being made.[26] We selected a cluster of 10 compounds with cell painting profiles most similar to the HDAC inhibitor vorinostat by UMAP analysis using the Ardigen phenAID JUMP-CP Data explorer (Figure 3A).[49] Using the structures of the compounds in this cluster as a focal graph query revealed histone deacetylases (HDACs 1,6,3 and 2) as the top-ranked targets (Figure 3B). This is based not only on the known targets of vorinostat, but also on the activities of other compounds in the cluster and their structurally related analogs (Figure 3C,D).

The focal graph visualization (Figure 3D) illustrates how relationships between the query compounds and HDAC family members were identified. Multiple independent lines of evidence connect the cluster of 10 compounds to HDAC1 (Figure 3C,D), including direct binding evidence and structural similarities to known HDAC inhibitors. This convergence of evidence from diverse sources provides strong support for HDAC inhibition as the shared mechanism of this set of compounds. This was summarized by a large language model analysis of the focal graph results (Figure 3E).

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

We also recently used focal graphs to analyze sets of genetic perturbations that were found to cluster together in a large cell painting dataset from the JUMP-Cell Painting Consortium, uncovering both expected and unexpected relationships between genes.[35] These results helped validate the dataset and uncovered new connections, identifying potential new drug targets that may play important roles in cancer, mitochondrial function, and nervous system development.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

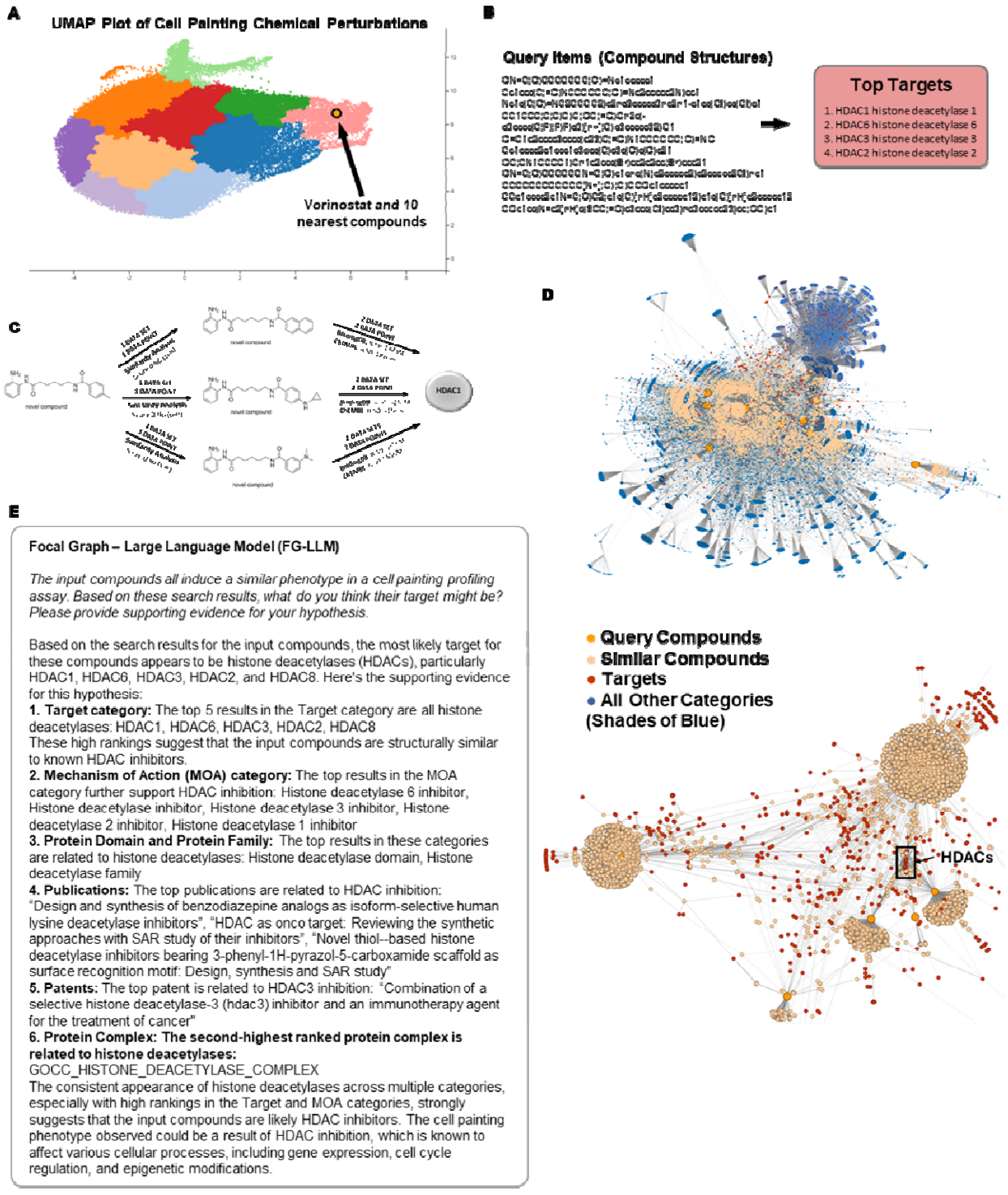


Figure 3. Focal graph analysis of a cell painting compound cluster. *A) Selection of a cluster of 10 compounds with cell painting profiles most similar to Vorinostat by UMAP analysis using the Ardigen phenAID JUMP-CP Data Explorer (https://phenaid.ardigen.com/jumpcpexplorer/). The x and y axes represent the first two dimensions of the UMAP projection, respectively. B) Top ranked targets from a focal graph search with the Vorinostat cell painting cluster, including Vorinostat itself. SMILES of the compounds in the cluster are shown. C) Selected focal graph evidence paths connecting a member of the compound series to HDAC1. D) The full vorinostat focal graph (top) and*

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

a subgraph (below) showing only query compounds, similar compounds, and targets. The location of a cluster of top ranked HDACs are highlighted. E) Large Language Model (LLM) analysis of the focal graph search which highlights HDAC inhibition as the best-supported target hypothesis for this compound cluster.[50–52] *Focal graphs were used in combination with Claude 3.5 Sonnet, a foundational LLM from Anthropic.*

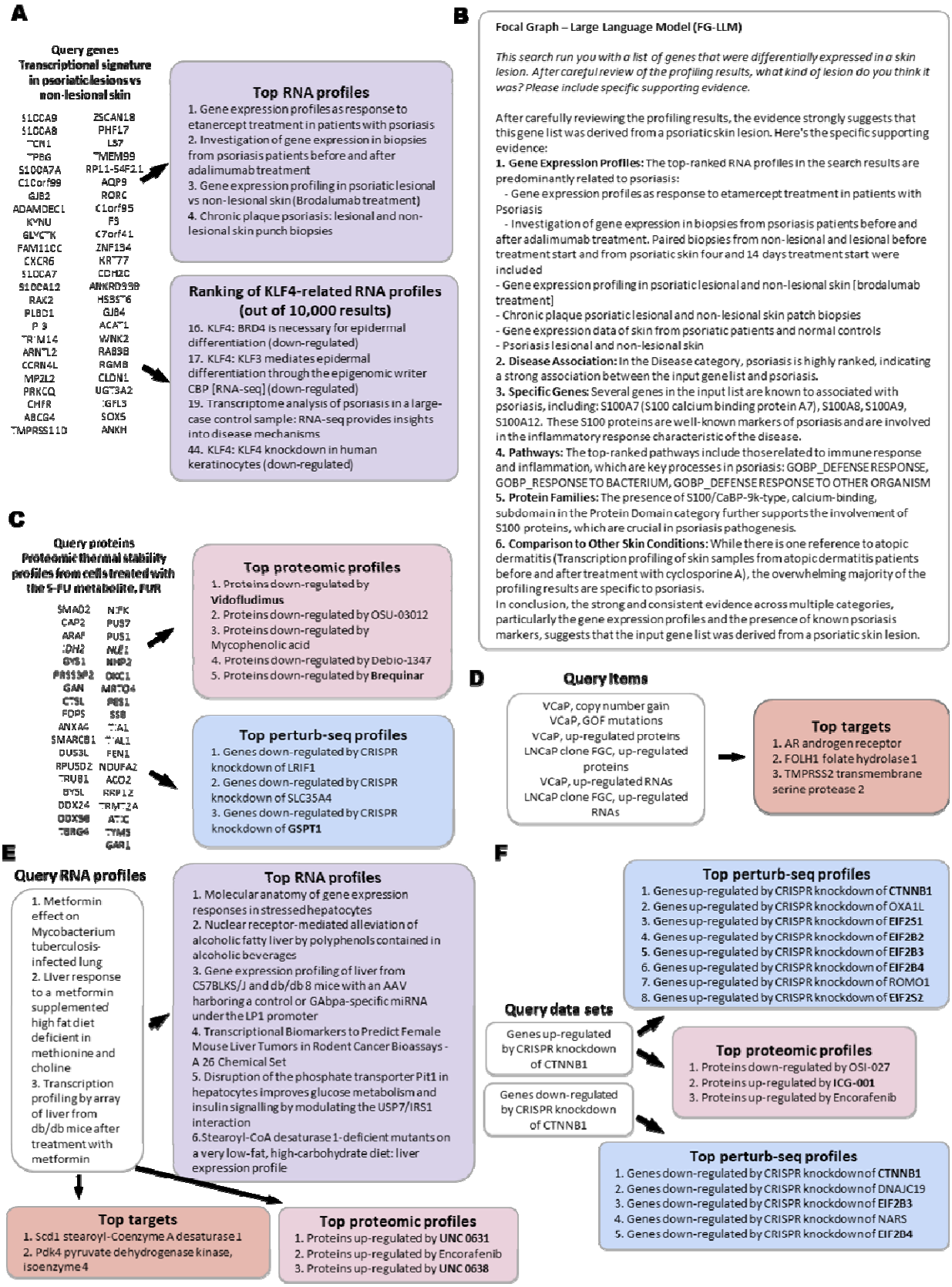
Multi-Omics Target Discovery

Gene expression profiling can provide an unbiased view of disease mechanisms, but extracting actionable insights from these complex datasets remains a challenge. We analyzed a transcriptional signature derived from psoriatic lesions compared to non-lesioned skin (Figure 4A).[53] The focal graph was generated using a list of differentially expressed genes, with no explicit reference to psoriasis. Nevertheless, the top-ranked results included multiple independent psoriasis gene expression studies (Figure 4A) providing strong validation of the signature's disease specificity. Large language model analysis of the search results (Figure 4B) systematically evaluated the evidence linking the signature to various skin conditions, confirming psoriasis as the best-supported diagnosis based purely on transcriptional changes.

Further analysis of the focal graph results revealed an intriguing connection to KLF4, with multiple KLF4-modulating profiles showing significant overlap with the disease signature (Figure 4A). KLF4, a transcription factor known to regulate epidermal differentiation,[54] appeared in four of the top 50 results out of nearly 10,000 total profiles (the ranking of each KLF4 RNA profile in the top 50 results is shown). This enrichment suggests that KLF4 modulation could potentially reverse the disease signature, making it a promising therapeutic target. The hypothesis is particularly compelling given KLF4's established role in maintaining skin barrier function and regulating inflammatory responses.[54,55]

This example illustrates how focal graphs can generate novel therapeutic hypotheses by integrating disease signatures with large-scale perturbational datasets. The approach not only validated the disease specificity of the signature but also identified a potential therapeutic target that might have been overlooked by traditional analysis methods. The transparent nature of the focal graph results allows direct evaluation of the evidence supporting the KLF4 hypothesis, from its effects on specific psoriasis-associated genes to its broader role in skin biology.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Figure 4. Focal graph analysis of omics profiles, discovery of novel pathway components & identification of biomarkers for precision medicine. *A) RNA profile results of a focal graph search using a transcriptional signature in psoriatic lesions versus non-lesional skin as a query. The top ranked RNA Profile results are independent gene expression studies of psoriasis. Profiles of genes after modulation by KLF4 are shown. These were ranked highly (rankings shown), including 4 results in the top 50 out of 9,889 total results. B) Large Language Model (LLM) analysis of the focal graph search which highlights psoriasis as the skin lesion type most strongly supported by the data, based solely on the differential gene expression profile.*[56–63] *C) Focal graph analysis of a CETSA proteomic thermal stability profile generated from cells treated with FUR, a 5-FU metabolite. The top proteomic profiles identified by the focal graph are shown. Two of the matching proteomic profiles are ones following treatment with DHODH inhibitors vidofludimus and brequinar. The top perturb seq profiles revealed by the focal graph included genes downregulated by CRISPR knockdown of GSPT1. D) Identification of biomarkers associated with prostate cancer using the molecular profiles of two prostate cancer cell lines VCaP and LNCaP as a query. The top targets identified were three of the most important genes linked to prostate cancer. E) 3 RNA profiles from mice treated with metformin were used as a query for a focal graph search. The top targets are strongly linked to glucose and fatty acid homeostasis. The top 100 genes from the RNA profile focal graph search were mapped to their human and rat orthologs and a signature representing all 3 species was run as another focal graph search. The results show links to RNA profiles related to glucose and fatty acid homeostasis. The top proteomic profiles show links to proteins up-regulated by small molecules UNC 0631 and UNC 0638, which are both inhibitors of EHMT2, which has been shown to be down-regulated by metformin. F) Discovery of novel components in the β-catenin pathway by focal graph analyses of genes up- or down-regulated after CRISPR knockdown of β-catenin. In each case the top Perturb seq profile results included multiple subunits of eukaryotic initiation factor 2 complex (eIF2). The top proteomic profiles contain a profile of proteins up-regulated by ICG-001, a known small molecule β-catenin inhibitor. Focal graphs were used in combination with Claude 3.5 Sonnet, a foundational LLM from Anthropic.*

Proteomic Profiling

Focal graphs can be used to analyze complex proteomic datasets. This is illustrated by analysis of proteomic thermal stability profiles generated from cells treated with the 5-FU metabolites FUR and FUDR and subjected to a cellular thermal shift assay, CETSA (Figure 4C).[64] 5-FU, a cornerstone of cancer chemotherapy for over 50 years, has long been known to inhibit thymidylate synthetase, but questions remain about its mechanism of action and strategies to improve its anti-cancer efficacy.[65,66] A focal graph analysis confirmed 5-FU's known effects on pyrimidine synthesis, and in addition, identified connections to proteomic profiles matching those following treatment with two DHODH inhibitors vidofludimus and brequinar (Figure 4C). DHODH inhibition has been reported to be synergistic with 5-FU treatment in cancer models.[67–69] DHODH is an enzyme in the pyrimidine biosynthesis pathway which produces uracil, providing a clear mechanism by which DHODH inhibitors may be synergistic with 5-FU treatment.

Intriguingly, the analysis revealed unexpected similarities between FUR/FUDR-induced proteomic changes and those caused by CRISPR knockdown of GSPT1 (Perturb seq profile in Figure 4C), suggesting interference with protein translation termination. This connection is

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

particularly noteworthy due to the recent emergence of GSPT1 as a novel clinical stage oncology target for which PROTACs have been developed.[70]

The CETSA example demonstrates the ability of the focal graph approach to weave together diverse data types, in this case connecting compound-proteomic interaction data (CETSA) to compound-induced proteomic protein level changes (MoA Atlas data)[71] and to transcriptomic RNA level changes after CRISPR knockdown (Perturb-Seq data).[72]

In summary, a focal graph analysis revealed additional data supporting possible synergy between 5-FU and DHODH inhibitors. It also identified a connection to GSPT1 knockdown, suggesting that the disruption of translational termination may be a part of 5-FU's mechanism and pointing to potential synergy between 5-FU and GSPT1 targeted therapeutics.

Precision Medicine

The development of robust patient stratification strategies such as the identification of biomarker signatures, remains a critical focus in precision oncology. Here we demonstrate how focal graphs can accelerate biomarker discovery by uncovering the molecular mechanisms that underlie a therapeutic response. A focal graph was queried with the molecular profiles of two prostate cancer cell lines (VCaP and LNCaP) that include RNA expression, protein levels, copy number alterations, and mutations.[73,74] The focal graph identified the most biologically significant genes associated with these cell lines as AR (androgen receptor), FOLH1 (PSMA), and TMPRSS2 (Figure 4D). These genes are arguably the most important, extensively validated and clinically relevant genes associated with prostate cancer. AR status guides treatment selection, FOLH1 expression enables PSMA-targeted imaging and therapy, and TMPRSS2 fusions serve as diagnostic and prognostic markers.[75–77]

This example illustrates a generalizable strategy for mechanism-based biomarker development. Lists of cell lines showing differential response to a therapeutic agent, can be analyzed by focal graphs to identify the molecular features that distinguish responders from non-responders. By analyzing large-scale cell line characterization datasets like CCLE and DepMap,[73,78,79] this approach can reveal not just correlative biomarkers but the underlying biological drivers of therapeutic sensitivity. The resulting mechanistic signatures can then be used to identify additional cell lines with similar molecular profiles, enabling rapid experimental validation of both the biomarker and its hypothesized mechanism of action.

The progression from response pattern to mechanistic understanding is particularly powerful for novel therapeutics where mechanisms are usually not immediately apparent.[80] The multi-omic nature of the analysis helps capture the full biological context of the drug response, while the transparent nature of focal graph analysis ensures that the proposed mechanisms are supported by clear experimental evidence. This mechanism-first approach could significantly accelerate the development of companion diagnostics by providing patient selection strategies with a solid mechanistic basis.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Safety Assessment

Focal graph analyses that are based on many types of queries, including chemical structures, multi-omics data, and cell line profiles, can be used to generate mechanism-of-toxicity hypotheses for safety assessment. Focal graphs starting from chemical structure were used to investigate the mechanism of drug-induced skin toxicity in cynomolgus monkeys.[14] The test compound was found to be structurally similar to compounds with a similar cell line proliferation profile to PI3K inhibitors. Many PI3K inhibitors are known to cause skin toxicity, suggesting a possible role for the PI3K pathway in the toxicity of the tested compound.[81]

Omics profiling, such as the psoriasis-derived RNA signature described previously, can be used to match a drug response to one or more profiles of a perturbagen with a described mechanism, which can then be used to identify potential mechanism(s) underlying the toxicity. A list of cell lines sensitive or resistant to a compound, such as those from a PRISM analysis,[82] can be used as a focal graph query to identify drug targets or mechanisms that may be driving toxicity. Many independent focal graph searches can be run in parallel, with mechanistic hypotheses assessed based on the strength of their supporting data from individual analyses, as well as agreement across analyses. This enables a multi-faceted, data-driven, comprehensive approach to safety assessment.

Biomarker Development

The development of pharmacodynamic biomarkers from gene expression signatures is another application of focal graphs. A focal graph analysis identified 100 mouse genes that were most commonly modulated across 3 independent transcriptional profiles of mice treated with metformin, a widely prescribed drug for the treatment of diabetes.[83–85] Notably, the top 2 genes in the signature, Scd1 and Pdk4 (Figure 4E), are known to be involved in glucose and fatty acid homeostasis. The 100 gene mouse signature, supplemented with the corresponding human and rat orthologs, was used in a subsequent focal graph search (Figure 4E), yielding top connections to other, independent transcriptional profiling studies related to the impacts of diet on hepatic function, including effects on insulin signaling and fatty liver. This suggests that this signature indeed captures the expected biology and that this strategy could be used as part of an effort to develop a pharmacodynamic biomarker for metformin response. Interestingly, two of the top three proteomic profiles overlapped the input signature by 6 of 100 genes. These are the profiles of UNC 0631 & UNC 0638, which both inhibit EHMT2, a histone lysine methyltransferase.[86,87] Metformin has been reported to down-regulate EHMT1 & EHMT2 expression at the RNA and protein levels.[88]

This example illustrates how focal graphs can be used to rapidly develop biomarkers using gene signatures as input. In this case the signatures were derived from drug-induced RNA profiles, but in principle they could come from many possible origins, including pathways, diseases, toxicities, and genetic perturbations, or from many possible data types, including transcriptomics, proteomics, genomics, metabolomics, or a mixture thereof.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Disease Indication Discovery

The discovery of novel disease indications for a target or a therapeutic is a further application of focal graphs in which disease processes based on gene signatures can be identified and characterized. When presented with a list of 50 up-regulated genes from a spatial, single-cell transcriptomics study of dystrophic muscle,[89] without any explicit mention of Duchenne Muscular Dystrophy (DMD), the integrated focal graph-LLM correctly identified DMD as the disease most strongly connected to the signature (Figure 5, S1). The system identified multiple independent RNA profiles specifically mentioning "mdx mice" - a well-established DMD model - and found studies directly comparing dystrophin-deficient and wild-type muscle tissue.

This example shows how focal graphs can be used to characterize disease processes in a relatively unbiased way, enabling connections to other data that can yield insights into the underlying disease mechanism and suggest new therapeutic areas.

Focal Graph-Integrated LLMs Outperform Non-Integrated LLMs

When the same gene list was analyzed by leading LLM models, including Claude 3.5, GPT-4, MedLM-Large, Co-Pilot Pro and Perplexity Pro, (Figure 5, S1), none of which include a focal graph analysis, none correctly identified DMD as the source of the signature.[32,90–93] Instead, these models variously attributed the gene expression pattern to cardiac fibrosis, atherosclerosis, Alzheimer's disease, or general inflammatory conditions. This discrepancy highlights a key advantage of focal graphs: the ability to ground LLMs in specific experimental data rather than relying solely on pattern matching from training data.

The DMD example demonstrates how focal graphs can enhance LLM capabilities by providing access to specialized datasets that may not be well-represented in general training data or which may not be well suited for current LLM training methods. The use of focal graphs as a connector enables the LLM's to synthesize and communicate complex biological insights while ensuring conclusions are supported by explicit experimental evidence. This combination of computational reasoning with empirical data validation represents a powerful new paradigm for biological data analysis. As the power and breadth of LLMs inevitably increases, so too should the power of combining them with other tools.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

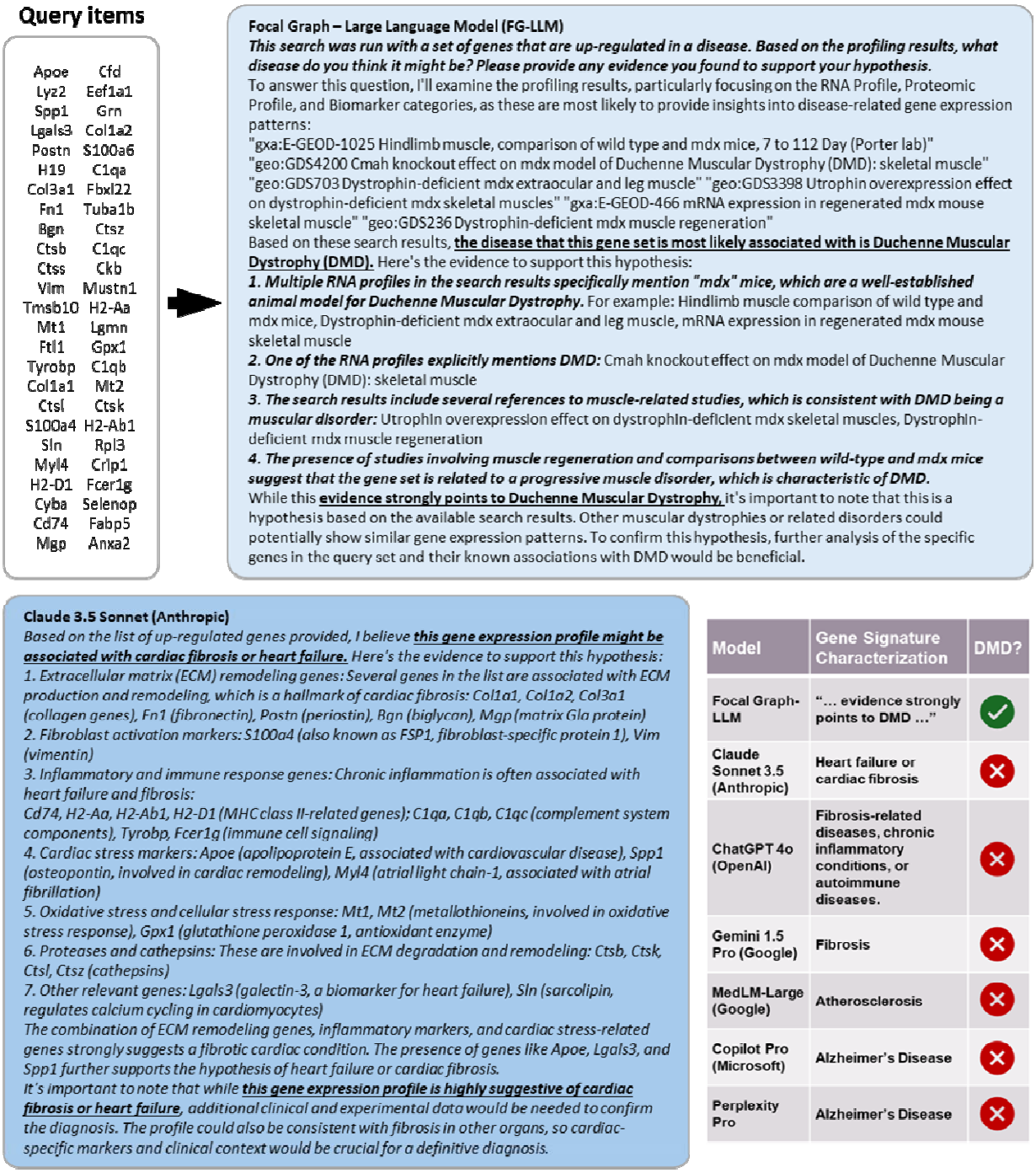


Figure 5. Identification of Duchenne Muscular Dystrophy from a gene signature by an LLM integrated with the focal graph. *Focal graph and LLM analysis of 50 genes upregulated in Duchenne Muscular Dystrophy (DMD) without any explicit mention of DMD. The LLM integrated with the focal graph correctly identified the disease based on the focal graph search results.[94–98] The same gene list was analyzed by leading LLM models alone, including Claude 3.5 Sonnet (Anthropic), ChatGPT-4o (OpenAI), MedLM-Large (Google), Gemini 1.5 Pro (Google), Co-Pilot Pro (Microsoft), and Perplexity Pro (Perplexity). None correctly identified DMD as the source of the signature attributing the gene expression pattern to cardiac fibrosis, atherosclerosis, Alzheimer's disease, or general inflammatory conditions. Focal graphs were*

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

used in combination with Claude 3.5 Sonnet, a foundational LLM from Anthropic. Underlined emphases added. See supplementary document S1 for unabridged transcripts of the LLM responses.

Systematic Discovery of Novel Pathway Components

The prediction of novel targets that could modulate disease relevant pathways is central to the development of new therapeutic strategies. The power of focal graphs to uncover novel pathway connections is demonstrated by focal graph analysis of Wnt pathway modulation. A focal graph analysis was run using mRNAs up- or down-regulated after CRISPR knockdown of β-catenin,[72] a central mediator of Wnt signaling, revealing unexpected connections to the translational machinery (Figure 4F). Multiple subunits of eukaryotic initiation factor 2 (EIF2) showed transcriptional responses that were highly correlated to β-catenin knockdown (Perturb seq datasets). This was found for multiple subunits of EIF2 and for both up- and down-regulated genes, suggesting that this is a robust connection.

The biological relevance of these findings was further reinforced by cross-modal validation, with strong correlations seen between β-catenin knockdown-induced transcripts and proteomic profiles from cells treated with ICG-001, a β-catenin inhibitor.[99] This orthogonal validation, spanning both genetic and pharmacological perturbations, and connecting transcriptomic to proteomic responses, suggests that this a *bona fide* biological connection.

This example illustrates how focal graphs can seed systematic pathway mapping efforts. Starting with established pathway components like β-catenin in Wnt signaling, the approach can iteratively expand to explore genetic perturbations, pharmacological modulation, protein-protein interactions, post-translational modifications, genetic dependencies, disease-associated mutations, and cell line profiling that could identify novel targets by which the Wnt pathway might be manipulated for therapeutic benefit.

This approach provides the basis for additional, automated focal graph searches and the foundation for autonomous AI-driven drug discovery, whereby each new connection can serve as a seed for additional focal graph searches. This allows systematic exploration of the pathway's molecular neighborhood, including cross-validation across multiple experimental approaches, model systems, and data types. Such an iterative strategy could be particularly valuable for identifying new or novel druggable nodes in traditionally challenging pathways in a highly automated, scalable fashion.

A Framework for Autonomous AI-Driven Drug Discovery

Wide Applicability of Focal Graphs

As we have shown, focal graphs can be applied to a wide range of drug discovery questions, and can be used to approach individual questions from many perspectives, using multiple data types and model systems. Additional applications in basic research include tool compound identification, model system discovery, metabolomics analysis, natural product target ID,

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

mechanism of resistance analyses, synthetic biology, microbiome analyses, and in studies of rare diseases. They can be applied across multiple stages of the drug discovery pipeline including in the design and characterization of screening libraries, hit to lead optimization, secondary pharmacology ("off-target analyses"), translational medicine, identification of combination therapies, and in the development of companion diagnostics. Further applications include drug repurposing, competitive intelligence, IP diligence & freedom to operate analyses, real world data analysis, veterinary medicine, quality assessment for cell-based therapies, and more. Focal graphs may have applicability across other industries as well, including agriculture, food science, environmental science, cosmetics and personal care, nutraceuticals, forensic science, biodefense and national security, waste management, and bioremediation.

Focal graph approaches have a long list of attributes that make them uniquely powerful and flexible, including transparency, concision, robustness to noise, compatibility with sparse data sets, the capacity to integrate algorithmic approaches including ML-based predictions, and the ability to mine extremely diverse sources of experimental data. These qualities should make them excellent agents for use in an LLM retrieval augmented generation (RAG) application, with focal graph searches being analogous (and complementary) to the internet searches commonly used in RAG-based applications such as Perplexity.[32]

Novelty and Bias

The knowledge graphs from which focal graphs are derived have biases that reflect their underlying data sources and the methods of their construction, so it is fair to ask what kinds of bias they may contain. Additionally, much of their information content may have been reported elsewhere, even if only in a database or a supplementary table, so likewise it is fair to ask to what extent they can enable novel discoveries.

Firstly, inclusion of less biased data sources, such as multi-omics data, should lead to less biased results. Secondly, it is important to note that when large scale data sources are reported in the scientific literature, only a small fraction of the data they generate is discussed explicitly in the text. The excess data points - which typically outnumber the text-mentioned data points by many orders of magnitude - are relegated to supplementary files or databases where their retrieval, analysis, and integration face the many hurdles we've already discussed. The result is that much of the data - that in theory has been reported - is in practice entirely unknown to the vast majority of researchers.

Finally, connecting dots across many disparate sources is a non-trivial but valuable exercise, which can lead to novel, potentially meaningful connections, as we have seen in many of the drug discovery examples presented here. It's impractical to ask a human to identify the hypothesis best supported by the data stored in the estimated ~6,000 existing biological databases,[3] or some appreciable fraction thereof, but this is precisely the kind of task that, in principle, can be easily accomplished by a focal graph analysis. In fact, judging by the focal graph results presented so far, such a query is likely to yield an abundance of interesting and novel results.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

Furthermore, it would be incorrect to assume that a focal graph analysis is unlikely to bear fruit when applied to obscure or understudied areas, such as rare diseases or novel chemistry. Much is conserved in evolution, and so a gene signature from a rare disease may closely resemble a well-studied biological process, or the knockdown of a well-understood gene in a large-scale omics experiment, yielding clues to its mechanism and suggesting possible therapeutic strategies. Novel chemistry can be mapped to the nearest informative chemical space. Even if the chemical similarity of its neighbors is only modest, integration with other data types may reveal coherent biology that is consistent with the known biological effects of the investigational compound. Alternatively, compound-derived biological profiles, such as multi-omics or cell line proliferation panels, can be used to glean insights into the compound's target or mechanism independent of its chemical structure.

In summary, although there is certainly bias in knowledge graphs, and by extension, focal graphs, they can nevertheless provide many opportunities for novel discovery; revealing what our "data knows" that human researchers may not.

Validation

Focal graphs are used here as *information retrieval systems*, much like internet search engines, that can surface information present in large knowledge graphs, and rank what they find by a measure of relevance. Formally speaking, they are not predictive models, although they can be informally used as such, and they could potentially be used to build true predictive models that could then be validated in the usual ways. Because the supporting data is conveyed in a transparent manner, their results can be evaluated in tremendous detail, and the confidence of their conclusions can be determined by domain experts.

Ultimately, validation of hypotheses generated by focal graphs comes down to expert evaluation of the data returned - its reproducibility, its quality, and its quantity - as well as the details of the sources and the methods used to generate it. Science is complex, and the details matter: Specific data points may support one conclusion but not another, and a conclusion that depends on a weak link may not be worth pursuing. Focal graph systems can mine vast amounts of data, surfacing what is most worth considering, generating rich data packages along the way, but in the end it is up to domain experts to judge the level of validation of any given hypothesis.

A Prototype Autonomous Target Discovery System

We have built a prototype focal graph-based RAG (FG-RAG) system and have conducted a preliminary analysis of its capabilities. The following prompt was used as a test case: "*Please plan and execute a research program to identify a novel oncology target in the Wnt pathway.*" An excerpted transcript of the response can be found in supplementary document 2. The FG-RAG proceeds to lay out a plan similar to the focal graph approach described previously, choosing to start with a β-catenin perturb seq profile as a first search, after which it nominates the EIF2 complex as an initial Wnt pathway target candidate, noting, "... *Several components of*

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

the eIF2 (eukaryotic initiation factor 2) complex appear in the top Perturb Seq results: [EIF2S1], [EIF2B2], [EIF2B3], [EIF2B4], and [EIF2B5]. This suggests a potential link between protein synthesis regulation and Wnt signaling. …" The FG-RAG continues its investigation, following up with 3 additional focal graph searches before summarizing its findings and proposing 3 potentially novel Wnt pathway targets: the eIF2 complex, OXA1L (a mitochondrial inner membrane protein), and SSB (Sjogren Syndrome Antigen B).

After this short initial investigation, which was run as a single process in less than 10 minutes, the FG-RAG could "decide" (or be manually instructed) to continue its perturb seq strategy with other Wnt pathway members, and/or move on to other focal graph search approaches using a variety of modalities, model systems, data types, etc. Such a system could continue its investigation indefinitely, amassing more and more validation as it goes, and increasing the quality, not just the quantity, of the targets it ultimately proposes. The extremely modest computational footprint means that truly massive runs of such a system are inherently feasible, and could result in the discovery of highly novel targets with vast amounts of experimental data to support them.

Towards More Powerful Autonomous Systems

Experimentally supported targets uncovered by focal graphs can be fed into other computational[100] or experimental[101] approaches, in a unidirectional or iterative way, to design novel therapeutics. These could be small molecules or other modalities, as determined by the scientific, health, and commercial considerations of the project.

The ability of FG-RAG systems to plan, execute, and interpret searches allows the creation of highly elaborate workflows, potentially involving iteration, multiple agent types, hierarchies, heuristics, and other complex decision making structures.[102,103] While individual components of drug discovery have been automated previously[21,104,105], the combination of focal graphs and LLMs allows the autonomous execution of extremely wide ranging research programs, while maintaining full transparency of their scientific rationale, research methods, and the data that supports their findings.

The potential scope of such a global, autonomous research effort is illustrated in Figure 6. The vast space of large-scale biomedical data sets can be explored systematically search by search and study by study, with a tremendous amount and diversity of data being brought to bear on any given research question. While a compelling answer cannot be guaranteed, the system can surface the answer best supported by available data, with subject area experts being the final arbiters of the assembled evidence and conclusions. The system can be further scaled to consider many research questions in parallel, highlighting cases where the data provides a compelling answer that appears to represent a significant scientific or medical breakthrough.

A gap analysis could lead to the identification of data sets whose incorporation would maximally increase the information content or novelty of future focal graph search results. Where existing data is insufficient, large-scale data generation campaigns could be launched. Combining an autonomous focal graph approach with a sufficiently powerful robotics-driven system could

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

further increase its power, allowing the automated, targeted generation of key data points that would enable further progress when a purely computational approach would otherwise reach its limit.

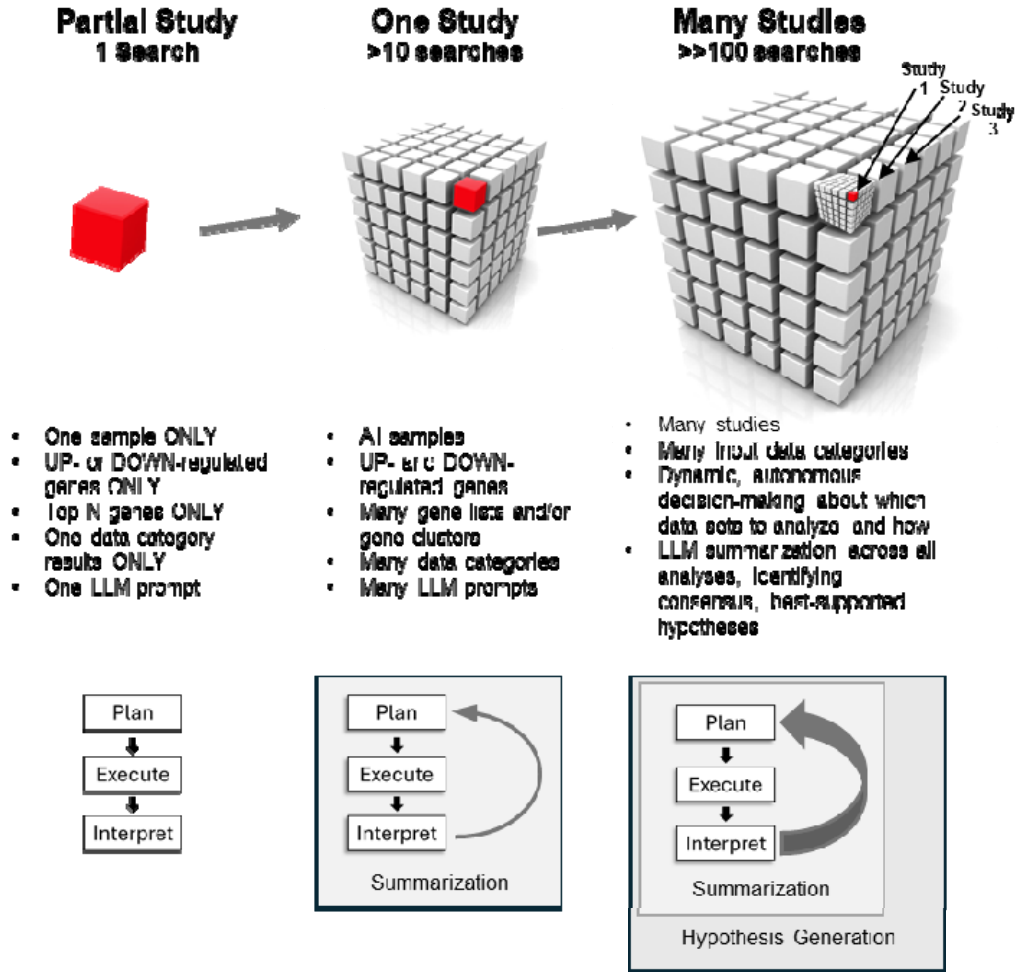


Figure 6. Autonomous planning and execution of research programs at massive scales. *The scale of an autonomous RAG-LLM research campaign could be vast, including systematic focal graph searches across large numbers of studies containing many modalities, model systems, data types, and approaches.*

Future Perspective

We have shown here that focal graph approaches are capable of generating novel, data-driven hypotheses from large scale data sets. Although the examples presented are limited, and would admittedly only represent the first steps in a real-world drug discovery program, they

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

nevertheless demonstrate discrete, verifiable progress - achieved by a highly scalable, transparent method that is capable of functioning with a high level of autonomy.

Furthermore, despite the cursory nature of these case studies, we were nevertheless able to uncover a number of interesting and plausible hypotheses backed by specific experimental data. There are an exceptionally wide range of potential focal graph applications, many of which will be the subject of future research reports. These projects can potentially be executed in tremendous depth, including across a breathtaking number of data types, modalities, and model systems. More research is needed to understand the best ways to create and utilize focal graphs, including how to optimally employ them as agents in LLM and RAG frameworks, potentially working side-by-side with other kinds of computational tools and even robotic systems.

It is likely that the answers to many questions of high scientific and medical importance remain hidden in the vast amounts of biomedical data that are largely inaccessible by current methods. This is due to a long list of reasons, including their dispersion across data repositories, their diversity, complexity, variable signal-to-noise, sparseness, scale, and inconsistent annotation, among others. We believe that the focal graph-based approaches described here can make a significant contribution. When combined with LLMs and other tools, focal graphs could transform our rich but underutilized data inheritance into a deeper understanding of the mechanisms underlying disease and into breakthrough therapies that could improve the human condition.

For too long, drug discovery efforts have been ineffective at using the vast amount of data produced by the global biomedical community. Each project generates and regenerates data anew, costing an unconscionable amount of time, resources, and ultimately, lives. It is past time that this effort became cumulative, with each new project truly standing on the shoulders of the giants that came before it. With the powerful tools we now have in hand - and with more appearing on the horizon - we may finally be on the verge of breaking through this tragic impasse. By learning from our historical data, rather than being doomed to forever repeat it, we may be able to make huge strides in our understanding of disease and in our relentless quest for cures.

Supplemental Information

Document S1. LLM Comparison
Document S2. FG-RAG Transcript (Excerpt)

Disclosures

All authors are employees or consultants of Plex Research, Inc. and may have real or optional ownership therein. Generative AI tools were used in the preparation of this manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

References

1. Ringel, M.S., Scannell, J.W., Baedeker, M., and Schulze, U. (2020). Breaking Eroom's Law. Nat. Rev. Drug Discov. *19*, 833–834. https://doi.org/10.1038/d41573-020-00059-3.
2. Scannell, J.W., Blanckley, A., Boldon, H., and Warrington, B. (2012). Diagnosing the decline in pharmaceutical R&D efficiency. Nat. Rev. Drug Discov. *11*, 191–200. https://doi.org/10.1038/nrd3681.
3. Ma, L., Zou, D., Liu, L., Shireen, H., Abbasi, A.A., Bateman, A., Xiao, J., Zhao, W., Bao, Y., and Zhang, Z. (2023). Database Commons: A Catalog of Worldwide Biological Databases. Genomics Proteomics Bioinformatics *21*, 1054–1058. https://doi.org/10.1016/j.gpb.2022.12.004.
4. Martínez-García, M., and Hernández-Lemus, E. (2022). Data Integration Challenges for Machine Learning in Precision Medicine. Front. Med. *8*. https://doi.org/10.3389/fmed.2021.784455.
5. Yang, X., Huang, K., Yang, D., Zhao, W., and Zhou, X. (2024). Biomedical Big Data Technologies, Applications, and Challenges for Precision Medicine: A Review. Glob. Chall. Hoboken NJ *8*, 2300163. https://doi.org/10.1002/gch2.202300163.
6. Hasselgren, C., and Oprea, T.I. (2024). Artificial Intelligence for Drug Discovery: Are We There Yet? Annu. Rev. Pharmacol. Toxicol. *64*, 527–550. https://doi.org/10.1146/annurev-pharmtox-040323-040828.
7. Santa Maria, J.P., Wang, Y., and Camargo, L.M. (2023). Perspective on the challenges and opportunities of accelerating drug discovery with artificial intelligence. Front. Bioinforma. *3*, 1121591. https://doi.org/10.3389/fbinf.2023.1121591.
8. Selinger, D.W., Shivashankar, V., Larbaoui, M., Mendelev, I., Steeves, M., Wilbert, M.L., Litster, S., Jenkins, J.L., and Marc, P. MoA Central: A Massively Orthogonal Search Engine for Mechanism of Action & Toxicity Studies. Preprint, https://doi.org/10.13140/RG.2.1.4278.4480 https://doi.org/10.13140/RG.2.1.4278.4480.
9. Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web.
10. Himmelstein, D.S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S.L., Hadley, D., Green, A., Khankhanian, P., and Baranzini, S.E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife *6*, e26726. https://doi.org/10.7554/eLife.26726.
11. Bonner, S., Barrett, I.P., Ye, C., Swiers, R., Engkvist, O., Bender, A., Hoyt, C.T., and Hamilton, W.L. (2022). A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. Brief. Bioinform. *23*, bbac404. https://doi.org/10.1093/bib/bbac404.
12. Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nat. Rev. Genet. *12*, 56–68. https://doi.org/10.1038/nrg2918.
13. Morselli Gysi, D., do Valle, Í., Zitnik, M., Ameli, A., Gan, X., Varol, O., Ghiassian, S.D., Patten, J.J., Davey, R.A., Loscalzo, J., et al. (2021). Network medicine framework for identifying drug-repurposing opportunities for COVID-19. Proc. Natl. Acad. Sci. U. S. A. *118*, e2025581118. https://doi.org/10.1073/pnas.2025581118.
14. Kikkawa, R. (2016). A Case of Drug-induced Cutaneous Toxicity Observed in Cynomolgus Monkeys. Toxicol. Pathol. *44*, 601–606. https://doi.org/10.1177/0192623316637710.
15. Hahn, M.W., and Kern, A.D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. Mol. Biol. Evol. *22*, 803–806. https://doi.org/10.1093/molbev/msi072.
16. Filippakopoulos, P., Qi, J., Picaud, S., Shen, Y., Smith, W.B., Fedorov, O., Morse, E.M., Keates, T., Hickman, T.T., Felletar, I., et al. (2010). Selective inhibition of BET bromodomains. Nature *468*, 1067–1073. https://doi.org/10.1038/nature09504.
17. Mendez, D., Gaulton, A., Bento, A.P., Chambers, J., De Veij, M., Félix, E., Magariños, M.P., Mosquera, J.F., Mutowo, P., Nowotka, M., et al. (2019). ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. *47*, D930–D940. https://doi.org/10.1093/nar/gky1075.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

18. Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., et al. (2023). PubChem 2023 update. Nucleic Acids Res. *51*, D1373–D1380. https://doi.org/10.1093/nar/gkac956.
19. Qian, C., Lai, C.-J., Bao, R., Wang, D.-G., Wang, J., Xu, G.-X., Atoyan, R., Qu, H., Yin, L., Samson, M., et al. (2012). Cancer network disruption by a single molecule inhibitor targeting both histone deacetylase activity and phosphatidylinositol 3-kinase signaling. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *18*, 4104–4113. https://doi.org/10.1158/1078-0432.CCR-12-0055.
20. Furet, P., Guagnano, V., Fairhurst, R.A., Imbach-Weese, P., Bruce, I., Knapp, M., Fritsch, C., Blasco, F., Blanz, J., Aichholz, R., et al. (2013). Discovery of NVP-BYL719 a potent and selective phosphatidylinositol-3 kinase alpha inhibitor selected for clinical evaluation. Bioorg. Med. Chem. Lett. *23*, 3741–3748. https://doi.org/10.1016/j.bmcl.2013.05.007.
21. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.
22. Walters, W.P., and Murcko, M. (2020). Assessing the impact of generative AI on medicinal chemistry. Nat. Biotechnol. *38*, 143–145. https://doi.org/10.1038/s41587-020-0418-2.
23. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., and Müller, K.-R. (2021). Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. Proc. IEEE *109*, 247–278. https://doi.org/10.1109/JPROC.2021.3060483.
24. Abdul Raheem, A.K., and Dhannoon, B.N. (2024). Comprehensive Review on Drug-target Interaction Prediction - Latest Developments and Overview. Curr. Drug Discov. Technol. *21*, e010923220652. https://doi.org/10.2174/1570163820666230901160043.
25. Xu, L., Ru, X., and Song, R. (2021). Application of Machine Learning for Drug–Target Interaction Prediction. Front. Genet. *12*. https://doi.org/10.3389/fgene.2021.680117.
26. Fredin Haslum, J., Lardeau, C.-H., Karlsson, J., Turkki, R., Leuchowius, K.-J., Smith, K., and Müllers, E. (2024). Cell Painting-based bioactivity prediction boosts high-throughput screening hit-rates and compound diversity. Nat. Commun. *15*, 3470. https://doi.org/10.1038/s41467-024-47171-1.
27. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F.A. (2020). Shortcut learning in deep neural networks. Nat. Mach. Intell. *2*, 665–673. https://doi.org/10.1038/s42256-020-00257-z.
28. Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J.K., Ceulemans, H., Clevert, D.-A., and Hochreiter, S. (2018). Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. Chem. Sci. *9*, 5441–5451. https://doi.org/10.1039/c8sc00148k.
29. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., and Sun, M. (2020). Graph neural networks: A review of methods and applications. AI Open *1*, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001.
30. Hwang, Y., Cornman, A.L., Kellogg, E.H., Ovchinnikov, S., and Girguis, P.R. (2024). Genomic language model predicts protein co-regulation and function. Nat. Commun. *15*, 2880. https://doi.org/10.1038/s41467-024-46947-9.
31. Sanabria, M., Hirsch, J., Joubert, P.M., and Poetsch, A.R. (2024). DNA language model GROVER learns sequence context in the human genome. Nat. Mach. Intell. *6*, 911–923. https://doi.org/10.1038/s42256-024-00872-0.
32. Perplexity Perplexity AI. https://www.perplexity.ai.
33. Sadeghi, S., Bui, A., Forooghi, A., Lu, J., and Ngom, A. (2024). Can large language models understand molecules? BMC Bioinformatics *25*, 225. https://doi.org/10.1186/s12859-024-05847-x.
34. Jablonka, K.M., Ai, Q., Al-Feghali, A., Badhwar, S., Bocarsly, J.D., Bran, A.M., Bringuier, S., Brinson, L.C., Choudhary, K., Circi, D., et al. (2023). 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. Digit. Discov. *2*, 1233–1250. https://doi.org/10.1039/d3dd00113j.
35. Chandrasekaran, S.N., Alix, E., Arevalo, J., Borowa, A., Byrne, P.J., Charles, W.G., Chen, Z.S., Cimini, B.A., Deng, B., Doench, J.G., et al. (2024). Morphological map of under- and over-expression

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

of genes in human cells. Preprint at bioRxiv, https://doi.org/10.1101/2024.12.02.624527 https://doi.org/10.1101/2024.12.02.624527.

36. Moffat, J.G., Rudolph, J., and Bailey, D. (2014). Phenotypic screening in cancer drug discovery - past, present and future. Nat. Rev. Drug Discov. *13*, 588–602. https://doi.org/10.1038/nrd4366.
37. Vincent, F., Nueda, A., Lee, J., Schenone, M., Prunotto, M., and Mercola, M. (2022). Phenotypic Drug Discovery: Recent successes, lessons learned and new directions. Nat. Rev. Drug Discov. *21*, 899–914. https://doi.org/10.1038/s41573-022-00472-w.
38. Schirle, M., and Jenkins, J.L. (2016). Identifying compound efficacy targets in phenotypic drug discovery. Drug Discov. Today *21*, 82–89. https://doi.org/10.1016/j.drudis.2015.08.001.
39. Wagner, B.K., and Schreiber, S.L. (2016). The power of sophisticated phenotypic screening and modern mechanism-of-action methods. Cell Chem. Biol. *23*, 3–9. https://doi.org/10.1016/j.chembiol.2015.11.008.
40. Calderón, F., Barros, D., Bueno, J.M., Coterón, J.M., Fernández, E., Gamo, F.J., Lavandera, J.L., León, M.L., Macdonald, S.J.F., Mallo, A., et al. (2011). An Invitation to Open Innovation in Malaria Drug Discovery: 47 Quality Starting Points from the TCAMS. ACS Med. Chem. Lett. *2*, 741–746. https://doi.org/10.1021/ml200135p.
41. Phillips, M.A., Lotharius, J., Marsh, K., White, J., Dayan, A., White, K.L., Njoroge, J.W., El Mazouni, F., Lao, Y., Kokkonda, S., et al. (2015). A long-duration dihydroorotate dehydrogenase inhibitor (DSM265) for prevention and treatment of malaria. Sci. Transl. Med. *7*, 296ra111. https://doi.org/10.1126/scitranslmed.aaa6645.
42. McCarthy, J.S., Lotharius, J., Rückle, T., Chalon, S., Phillips, M.A., Elliott, S., Sekuloski, S., Griffin, P., Ng, C.L., Fidock, D.A., et al. (2017). Safety, tolerability, pharmacokinetics, and activity of the novel long-acting antimalarial DSM265: a two-part first-in-human phase 1a/1b randomised study. Lancet Infect. Dis. *17*, 626–635. https://doi.org/10.1016/S1473-3099(17)30171-8.
43. Abdullah, I., Chee, C.F., Lee, Y.-K., Thunuguntla, S.S.R., Satish Reddy, K., Nellore, K., Antony, T., Verma, J., Mun, K.W., Othman, S., et al. (2015). Benzimidazole derivatives as potential dual inhibitors for PARP-1 and DHODH. Bioorg. Med. Chem. *23*, 4669–4680. https://doi.org/10.1016/j.bmc.2015.05.051.
44. Palmer, M.J., Deng, X., Watts, S., Krilov, G., Gerasyuto, A., Kokkonda, S., El Mazouni, F., White, J., White, K.L., Striepen, J., et al. (2021). Potent Antimalarials with Development Potential Identified by Structure-Guided Computational Optimization of a Pyrrole-Based Dihydroorotate Dehydrogenase Inhibitor Series. J. Med. Chem. *64*, 6085–6136. https://doi.org/10.1021/acs.jmedchem.1c00173.
45. Inaoka, D.K., Iida, M., Hashimoto, S., Tabuchi, T., Kuranaga, T., Balogun, E.O., Honma, T., Tanaka, A., Harada, S., Nara, T., et al. (2017). Design and synthesis of potent substrate-based inhibitors of the Trypanosoma cruzi dihydroorotate dehydrogenase. Bioorg. Med. Chem. *25*, 1465–1470. https://doi.org/10.1016/j.bmc.2017.01.009.
46. Bray, M.-A., Singh, S., Han, H., Davis, C.T., Borgeson, B., Hartland, C., Kost-Alimova, M., Gustafsdottir, S.M., Gibson, C.C., and Carpenter, A.E. (2016). Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat. Protoc. *11*, 1757–1774. https://doi.org/10.1038/nprot.2016.105.
47. Seal, S., Trapotsi, M.-A., Spjuth, O., Singh, S., Carreras-Puigvert, J., Greene, N., Bender, A., and Carpenter, A.E. (2024). Cell Painting: a decade of discovery and innovation in cellular imaging. Nat. Methods, 1–15. https://doi.org/10.1038/s41592-024-02528-8.
48. Chandrasekaran, S.N., Ceulemans, H., Boyd, J.D., and Carpenter, A.E. (2021). Image-based profiling for drug discovery: due for a machine-learning upgrade? Nat. Rev. Drug Discov. *20*, 145–159. https://doi.org/10.1038/s41573-020-00117-w.
49. phenAID https://phenaid.ardigen.com/jumpcpexplorer/.
50. Reddy, D.R., Ballante, F., Zhou, N.J., and Marshall, G.R. (2017). Design and synthesis of benzodiazepine analogs as isoform-selective human lysine deacetylase inhibitors. Eur. J. Med. Chem. *127*, 531–553. https://doi.org/10.1016/j.ejmech.2016.12.032.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

51. Sangwan, R., Rajan, R., and Mandal, P.K. (2018). HDAC as onco target: Reviewing the synthetic approaches with SAR study of their inhibitors. Eur. J. Med. Chem. *158*, 620–706. https://doi.org/10.1016/j.ejmech.2018.08.073.
52. Wen, J., Niu, Q., Liu, J., Bao, Y., Yang, J., Luan, S., Fan, Y., Liu, D., and Zhao, L. (2016). Novel thiol-based histone deacetylase inhibitors bearing 3-phenyl-1H-pyrazole-5-carboxamide scaffold as surface recognition motif: Design, synthesis and SAR study. Bioorg. Med. Chem. Lett. *26*, 375–379. https://doi.org/10.1016/j.bmcl.2015.12.007.
53. Di Meglio, P., Duarte, J.H., Ahlfors, H., Owens, N.D.L., Li, Y., Villanova, F., Tosi, I., Hirota, K., Nestle, F.O., Mrowietz, U., et al. (2014). Activation of the aryl hydrocarbon receptor dampens the severity of inflammatory skin conditions. Immunity *40*, 989–1001. https://doi.org/10.1016/j.immuni.2014.04.019.
54. Segre, J.A., Bauer, C., and Fuchs, E. (1999). Klf4 is a transcription factor required for establishing the barrier function of the skin. Nat. Genet. *22*, 356–360. https://doi.org/10.1038/11926.
55. Liang, Y., Zhao, J., Dai, T., Li, X., Chen, L., He, Z., Guo, M., Zhao, J., and Xu, L. (2024). A review of KLF4 and inflammatory disease: Current status and future perspective. Pharmacol. Res. *207*, 107345. https://doi.org/10.1016/j.phrs.2024.107345.
56. Correa da Rosa, J., Kim, J., Tian, S., Tomalin, L.E., Krueger, J.G., and Suárez-Fariñas, M. (2017). Shrinking the Psoriasis Assessment Gap: Early Gene-Expression Profiling Accurately Predicts Response to Long-Term Treatment. J. Invest. Dermatol. *137*, 305–312. https://doi.org/10.1016/j.jid.2016.09.015.
57. Russell, C.B., Rand, H., Bigler, J., Kerkof, K., Timour, M., Bautista, E., Krueger, J.G., Salinger, D.H., Welcher, A.A., and Martin, D.A. (2014). Gene expression profiles normalized in psoriatic skin by treatment with brodalumab, a human anti-IL-17 receptor monoclonal antibody. J. Immunol. Baltim. Md 1950 *192*, 3828–3836. https://doi.org/10.4049/jimmunol.1301737.
58. Swindell, W.R., Xing, X., Stuart, P.E., Chen, C.S., Aphale, A., Nair, R.P., Voorhees, J.J., Elder, J.T., Johnston, A., and Gudjonsson, J.E. (2012). Heterogeneity of inflammatory and cytokine networks in chronic plaque psoriasis. PloS One *7*, e34594. https://doi.org/10.1371/journal.pone.0034594.
59. Nair, R.P., Duffin, K.C., Helms, C., Ding, J., Stuart, P.E., Goldgar, D., Gudjonsson, J.E., Li, Y., Tejasvi, T., Feng, B.-J., et al. (2009). Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. Nat. Genet. *41*, 199–204. https://doi.org/10.1038/ng.311.
60. Swindell, W.R., Johnston, A., Carbajal, S., Han, G., Wohn, C., Lu, J., Xing, X., Nair, R.P., Voorhees, J.J., Elder, J.T., et al. (2011). Genome-wide expression profiling of five mouse models identifies similarities and differences with human psoriasis. PloS One *6*, e18266. https://doi.org/10.1371/journal.pone.0018266.
61. Ding, J., Gudjonsson, J.E., Liang, L., Stuart, P.E., Li, Y., Chen, W., Weichenthal, M., Ellinghaus, E., Franke, A., Cookson, W., et al. (2010). Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. Am. J. Hum. Genet. *87*, 779–789. https://doi.org/10.1016/j.ajhg.2010.10.024.
62. Khattri, S., Shemer, A., Rozenblit, M., Dhingra, N., Czarnowicki, T., Finney, R., Gilleaudeau, P., Sullivan-Whalen, M., Zheng, X., Xu, H., et al. (2014). Cyclosporine in patients with atopic dermatitis modulates activated inflammatory pathways and reverses epidermal pathology. J. Allergy Clin. Immunol. *133*, 1626–1634. https://doi.org/10.1016/j.jaci.2014.03.003.
63. E-MTAB-5262 Investig. Gene Expr. Biopsies Psoriasis Patients Adalimumab Treat. Paired Biopsies Non-Lesional Lesional Treat. Start Psoriatic Skin Four 14 Days Treat. Start Were Incl. https://www.ebi.ac.uk/gxa/experiments/E-MTAB-5262/Results.
64. Liang, Y.Y., Bacanu, S., Sreekumar, L., Ramos, A.D., Dai, L., Michaelis, M., Cinatl, J., Seki, T., Cao, Y., Coffill, C.R., et al. (2022). CETSA interaction proteomics define specific RNA-modification pathways as key components of fluorouracil-based cancer drug cytotoxicity. Cell Chem. Biol. *29*, 572-585.e8. https://doi.org/10.1016/j.chembiol.2021.06.007.
65. A Fifty-Year-Old Cancer Drug Doesn't Do What You Think https://www.science.org/content/blog-post/fifty-year-old-cancer-drug-doesn-t-do-what-you-think.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

66. Longley, D.B., Harkin, D.P., and Johnston, P.G. (2003). 5-fluorouracil: mechanisms of action and clinical strategies. Nat. Rev. Cancer *3*, 330–338. https://doi.org/10.1038/nrc1074.
67. Mohamad Fairus, A.K., Choudhary, B., Hosahalli, S., Kavitha, N., and Shatrah, O. (2017). Dihydroorotate dehydrogenase (DHODH) inhibitors affect ATP depletion, endogenous ROS and mediate S-phase arrest in breast cancer cells. Biochimie *135*, 154–163. https://doi.org/10.1016/j.biochi.2017.02.003.
68. Amalia, E., Diantini, A., Endang Prabandari, E., Waluyo, D., and Subarnas, A. (2022). Caffeic Acid Phenethyl Ester as a DHODH Inhibitor and Its Synergistic Anticancer Properties in Combination with 5-Fluorouracil in a Breast Cancer Cell Line. J. Exp. Pharmacol. *14*, 243–253. https://doi.org/10.2147/JEP.S365159.
69. Pizzorno, G., Wiegand, R.A., Lentz, S.K., and Handschumacher, R.E. (1992). Brequinar potentiates 5-fluorouracil antitumor activity in a murine model colon 38 tumor by tissue-specific modulation of uridine nucleotide pools. Cancer Res. *52*, 1660–1665.
70. Zhang, D., Lin, P., and Lin, J. (2024). Molecular glues targeting GSPT1 in cancers: A potent therapy. Bioorganic Chem. *143*, 107000. https://doi.org/10.1016/j.bioorg.2023.107000.
71. Mitchell, D.C., Kuljanin, M., Li, J., Van Vranken, J.G., Bulloch, N., Schweppe, D.K., Huttlin, E.L., and Gygi, S.P. (2023). A proteome-wide atlas of drug mechanism of action. Nat. Biotechnol. *41*, 845–857. https://doi.org/10.1038/s41587-022-01539-0.
72. Replogle, J.M., Saunders, R.A., Pogson, A.N., Hussmann, J.A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E.J., Adelman, K., Lithwick-Yanai, G., et al. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell *185*, 2559-2575.e28. https://doi.org/10.1016/j.cell.2022.05.013.
73. Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M., et al. (2017). Defining a Cancer Dependency Map. Cell *170*, 564-576.e16. https://doi.org/10.1016/j.cell.2017.06.010.
74. DepMap 24Q2 Public (2024). https://plus.figshare.com/articles/dataset/DepMap_24Q2_Public/25880521/1.
75. Armstrong, A.J., Eisenberger, M.A., Halabi, S., Oudard, S., Nanus, D.M., Petrylak, D.P., Sartor, A.O., and Scher, H.I. (2012). Biomarkers in the management and treatment of men with metastatic castration-resistant prostate cancer. Eur. Urol. *61*, 549–559. https://doi.org/10.1016/j.eururo.2011.11.009.
76. Kumar, A.A. (2024). Prostate cancer genotyping for risk stratification and precision treatment. Curr. Urol. *18*, 87–97. https://doi.org/10.1097/CU9.0000000000000222.
77. Tayara, O., Poletajew, S., Malewski, W., Kunikowska, J., Pełka, K., Kryst, P., and Nyk, Ł. (2024). Prostate-Specific Membrane Antigen Expression in Patients with Primary Prostate Cancer: Diagnostic and Prognostic Value in Positron Emission Tomography-Prostate-Specific Membrane Antigen. Curr. Oncol. Tor. Ont *31*, 4165–4177. https://doi.org/10.3390/curroncol31080311.
78. Ghandi, M., Huang, F.W., Jané-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature *569*, 503–508. https://doi.org/10.1038/s41586-019-1186-3.
79. Nusinow, D.P., Szpyt, J., Ghandi, M., Rose, C.M., McDonald, E.R., Kalocsay, M., Jané-Valbuena, J., Gelfand, E., Schweppe, D.K., Jedrychowski, M., et al. (2020). Quantitative Proteomics of the Cancer Cell Line Encyclopedia. Cell *180*, 387-402.e16. https://doi.org/10.1016/j.cell.2019.12.023.
80. Lara Gongora, A.B., Carvalho Oliveira, L.J., and Jardim, D.L. (2020). Impact of the biomarker enrichment strategy in drug development. Expert Rev. Mol. Diagn. *20*, 611–618. https://doi.org/10.1080/14737159.2020.1711734.
81. Jfri, A., Meltzer, R., Mostaghimi, A., LeBoeuf, N., and Guggina, L. (2022). Incidence of Cutaneous Adverse Events With Phosphoinositide 3-Kinase Inhibitors as Adjuvant Therapy in Patients With Cancer: A Systematic Review and Meta-analysis. JAMA Oncol. *8*, 1635–1643. https://doi.org/10.1001/jamaoncol.2022.4327.

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

82. Yu, C., Mannan, A.M., Yvone, G.M., Ross, K.N., Zhang, Y.-L., Marton, M.A., Taylor, B.R., Crenshaw, A., Gould, J.Z., Tamayo, P., et al. (2016). High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. Nat. Biotechnol. *34*, 419–423. https://doi.org/10.1038/nbt.3460.
83. Heishi, M., Ichihara, J., Teramoto, R., Itakura, Y., Hayashi, K., Ishikawa, H., Gomi, H., Sakai, J., Kanaoka, M., Taiji, M., et al. (2006). Global gene expression analysis in liver of obese diabetic db/db mice treated with metformin. Diabetologia *49*, 1647–1655. https://doi.org/10.1007/s00125-006-0271-y.
84. Singhal, A., Jie, L., Kumar, P., Hong, G.S., Leow, M.K.-S., Paleja, B., Tsenova, L., Kurepina, N., Chen, J., Zolezzi, F., et al. (2014). Metformin as adjunct antituberculosis therapy. Sci. Transl. Med. *6*, 263ra159. https://doi.org/10.1126/scitranslmed.3009885.
85. Kita, Y., Takamura, T., Misu, H., Ota, T., Kurita, S., Takeshita, Y., Uno, M., Matsuzawa-Nagata, N., Kato, K.-I., Ando, H., et al. (2012). Metformin prevents and reverses inflammation in a non-diabetic mouse model of nonalcoholic steatohepatitis. PloS One *7*, e43056. https://doi.org/10.1371/journal.pone.0043056.
86. Vedadi, M., Barsyte-Lovejoy, D., Liu, F., Rival-Gervier, S., Allali-Hassani, A., Labrie, V., Wigle, T.J., Dimaggio, P.A., Wasney, G.A., Siarheyeva, A., et al. (2011). A chemical probe selectively inhibits G9a and GLP methyltransferase activity in cells. Nat. Chem. Biol. *7*, 566–574. https://doi.org/10.1038/nchembio.599.
87. Liu, F., Barsyte-Lovejoy, D., Allali-Hassani, A., He, Y., Herold, J.M., Chen, X., Yates, C.M., Frye, S.V., Brown, P.J., Huang, J., et al. (2011). Optimization of cellular activity of G9a inhibitors 7-aminoalkoxy-quinazolines. J. Med. Chem. *54*, 6139–6150. https://doi.org/10.1021/jm200903z.
88. Kim, D., Kim, Y., Lee, B.B., Cho, E.Y., Han, J., Shim, Y.M., and Kim, D.-H. (2021). Metformin Reduces Histone H3K4me3 at the Promoter Regions of Positive Cell Cycle Regulatory Genes in Lung Cancer Cells. Cancers *13*, 739. https://doi.org/10.3390/cancers13040739.
89. Stec, M.J., Su, Q., Adler, C., Zhang, L., Golann, D.R., Khan, N.P., Panagis, L., Villalta, S.A., Ni, M., Wei, Y., et al. (2023). A cellular and molecular spatial atlas of dystrophic muscle. Proc. Natl. Acad. Sci. U. S. A. *120*, e2221249120. https://doi.org/10.1073/pnas.2221249120.
90. Claude. https://claude.ai.
91. ChatGPT https://chatgpt.com.
92. Microsoft Copilot: Your AI companion Microsoft Copilot Your AI Companion. https://copilot.microsoft.com.
93. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. (2023). Large language models encode clinical knowledge. Nature *620*, 172–180. https://doi.org/10.1038/s41586-023-06291-2.
94. Chandrasekharan, K., Yoon, J.H., Xu, Y., deVries, S., Camboni, M., Janssen, P.M.L., Varki, A., and Martin, P.T. (2010). A human-specific deletion in mouse Cmah increases disease severity in the mdx model of Duchenne muscular dystrophy. Sci. Transl. Med. *2*, 42ra54. https://doi.org/10.1126/scitranslmed.3000692.
95. Porter, J.D., Khanna, S., Kaminski, H.J., Rao, J.S., Merriam, A.P., Richmonds, C.R., Leahy, P., Li, J., Guo, W., and Andrade, F.H. (2002). A chronic inflammatory response dominates the skeletal muscle molecular signature in dystrophin-deficient mdx mice. Hum. Mol. Genet. *11*, 263–272. https://doi.org/10.1093/hmg/11.3.263.
96. Baban, D., and Davies, K.E. (2008). Microarray analysis of mdx mice expressing high levels of utrophin: therapeutic implications for dystrophin deficiency. Neuromuscul. Disord. NMD *18*, 239–247. https://doi.org/10.1016/j.nmd.2007.11.011.
97. Porter, J.D., Merriam, A.P., Leahy, P., Gong, B., and Khanna, S. (2003). Dissection of temporal gene expression signatures of affected and spared muscle groups in dystrophin-deficient (mdx) mice. Hum. Mol. Genet. *12*, 1813–1821. https://doi.org/10.1093/hmg/ddg197.
98. Tseng, B.S., Zhao, P., Pattison, J.S., Gordon, S.E., Granchelli, J.A., Madsen, R.W., Folk, L.C., Hoffman, E.P., and Booth, F.W. (2002). Regenerated mdx mouse skeletal muscle shows differential

bioRxiv preprint doi: https://doi.org/10.1101/2024.12.17.629024; this version posted December 20, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license.

mRNA expression. J. Appl. Physiol. Bethesda Md 1985 *93*, 537–545. https://doi.org/10.1152/japplphysiol.00202.2002.

99. Emami, K.H., Nguyen, C., Ma, H., Kim, D.H., Jeong, K.W., Eguchi, M., Moon, R.T., Teo, J.-L., Kim, H.Y., Moon, S.H., et al. (2004). A small molecule inhibitor of beta-catenin/CREB-binding protein transcription [corrected]. Proc. Natl. Acad. Sci. U. S. A. *101*, 12682–12687. https://doi.org/10.1073/pnas.0404875101.
100. Atomwise AIMS Program (2024). AI is a viable alternative to high throughput screening: a 318-target study. Sci. Rep. *14*, 7526. https://doi.org/10.1038/s41598-024-54655-z.
101. Gironda-Martínez, A., Donckele, E.J., Samain, F., and Neri, D. (2021). DNA-Encoded Chemical Libraries: A Comprehensive Review with Successful Stories and Future Challenges. ACS Pharmacol. Transl. Sci. *4*, 1265–1279. https://doi.org/10.1021/acsptsci.1c00118.
102. Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J.R., Ektefaie, Y., Kondic, J., and Zitnik, M. (2024). Empowering biomedical discovery with AI agents. Cell *187*, 6125–6151. https://doi.org/10.1016/j.cell.2024.09.022.
103. Zhou, J., Zhang, B., Li, G., Chen, X., Li, H., Xu, X., Chen, S., He, W., Xu, C., Liu, L., et al. (2024). An AI Agent for Fully Automated Multi-Omic Analyses. Adv. Sci. Weinh. Baden-Wurtt. Ger. *11*, e2407094. https://doi.org/10.1002/advs.202407094.
104. Zheng, Y., Koh, H.Y., Yang, M., Li, L., May, L.T., Webb, G.I., Pan, S., and Church, G. (2024). Large Language Models in Drug Discovery and Development: From Disease Mechanisms to Clinical Trials. Preprint at arxiv.org, https://doi.org/10.48550/arXiv.2409.04481 https://doi.org/10.48550/arXiv.2409.04481.
105. Huang, K., Fu, T., Glass, L.M., Zitnik, M., Xiao, C., and Sun, J. (2021). DeepPurpose: a deep learning library for drug–target interaction prediction. Bioinformatics *36*, 5545–5547. https://doi.org/10.1093/bioinformatics/btaa1005.