**Introduction:** AI scientists are systems of LLM-based agents that are trained to use tools, conduct multi-step reasoning, collaborate with each other and humans, and interact with the world in order to conduct scientific research. These systems can accelerate scientific discovery by identifying overlooked insights in the literature, automating repetitive tasks, and creating cheaper and more efficient computational research pipelines. These benefits are particularly impactful in the chemical and molecular sciences, where accelerating research in protein, molecular, and materials design could translate to medical breakthroughs and new industrial applications.

Current AI scientists have shown promise in autonomously designing functional proteins [1] and conducting bioinformatics analyses [2]. In the Zitnik Lab, I helped develop TxAgent, a state-of-the-art AI scientist for reasoning about drugs, personalized treatments, and adverse event prediction [3]. However, a major gap in the field is that AI scientists lack the ability to reason about visual and geometric data, such as protein and molecular structures or experimental visualizations (plots, graphs, and images of the laboratory). This prevents AI scientists from autonomously performing modern chemical research.

While existing general-purpose multimodal models like Gemini 2.5 or GPT-4o could be applied to visual and geometric reasoning tasks in chemical and biochemical research, they struggle to adapt to visualization and geometric challenges specific to the molecular domain [4]. For example, a protein has a complex 3-D structure that cannot be easily encoded in a 2D image or 1D textual description, preventing general-purpose multimodal models like Gemini 2.5 from understanding the protein. Other chemical data types like spectra graphs (which capture how a molecule interacts with light) are domain-specific visualizations that a general-purpose image encoder may struggle to encode.

In response, people have constructed domain-specific multimodal models that reason about specific chemical data modalities. For example, Cui *et al.* use a specialized encoder for crystals (a class of solid chemical compounds that includes diamonds and crystalline silicon) [5]. With this "crystal encoder," they build a multimodal model that can reason about structures and textual descriptions of crystals [5]. In this proposal, I extend this idea to a more modalities, creating a general-purpose, multimodal framework for molecular perception and reasoning that can be paired with tool-using AI scientist systems (*e.g.*, via the AI scientist TxAgent and its toolbox, ToolUniverse, that I helped develop in the Zitnik Lab) to construct true AI chemists, material scientists, and protein designers.

**Related Work:** Existing multimodal models for chemical and molecular reasoning tend to focus on one or a select few modalities. For example, ChemVLM [6] uses text and vision transformers to reason about only small molecules, while $L^2M^3OF$ [5] is a multimodal model that only reasons about metal-organic frameworks, a subcategory of crystals. As a result, we don't yet have a general-purpose multimodal model that can encode a diverse array of chemical data types—from protein structures and drug molecules to laboratory images—that informs its reasoning and textual output.

**Proposal:** To allow AI scientists to perceive and manipulate chemical data, I will build MolPerceptor, a multimodal language model that reasons about molecular data (small molecules, proteins, and crystals) and laboratory data (plots and tables) via specialized encoders for each data type. I hypothesize this will result in improved multimodal perception capabilities. MolPerceptor can then be integrated into an AI scientist system so the AI scientist can reason more accurately about the chemical and molecular world.

**Aim I: Building MolPerceptor's Architecture.** Using domain-specific encoders, I can produce vector embedding representations of complex chemical data types that can be integrated into MolPerception's base language model. Therefore, for each data modality (protein structures, crystal structures, spectra graphs or other plots, and images of laboratory equipment), I will identify a state-of-the-art encoder that takes in the data modality and outputs a vector embedding. Two examples of these encoders are:

- ATOMICA is a geometric AI model that takes in a 3-D graph of two interacting molecules (*e.g.*, a protein and a drug) and outputs a vector embedding capturing the interaction information.
- CrystalFormer is a transformer model that takes in matrix representations of the atom types, positions, and orientations of the crystal and outputs a vector embedding of this crystal structure.

After applying encoders to each data modality, I must integrate the output embeddings with a base language model. To do this, I will apply cross-attention between the base model's text embeddings and the embeddings for each non-text modality.

**Aim II: Training MolPerceptor.** To train this multimodal architecture, I first must align the various encoders. I will use contrastive losses (*e.g.*, the infoNCE loss) to align vector embeddings of the same entity across different modalities. For example, a molecule like ethanol (alcohol) can be represented as text using SMILES notation, as a graph of atoms and bonds, or as an NMR spectrum diagram showing the arrangement of different parts of the molecule. Then, our text, graph, and spectra encoders would yield three different vector embeddings for ethanol that must be brought closer together in embedding space.

After the encoders are aligned, I will then train the multimodal model. To do this, I must collect a corpus of multimodal chemical questions and answers that feature chemical structures, images, and text. I will use a two-stage data collection and augmentation pipeline. In Stage 1, I will use recent vision-language models designed to collect multimodal data (images, reactions, and text) from chemical literature [7]. Using these models, I will build a corpus of multimodal information that spans proteins, small molecules, and crystals. In Stage 2, I will use the multi-agent question generation pipelines from the TxAgent project to generate questions and answers from our multimodal information corpus. These pipelines will use LLM agents to extract useful information from our corpus, divide this information into question-answer pairs, and reframe the questions in different ways to ensure robustness in reasoning. For example, a question may show a molecule's geometric structure, while the reframed question shows the structure encoded in text form as a SMILES string. This will reduce the performance discrepancies across different modalities seen in current multimodal models for chemical reasoning [4].

**Aim III: Augmenting AI scientists with MolPerceptor.** MolPerceptor is an AI model that can reason about diverse chemical and molecular data modalities. Then, I will integrate it into existing AI scientists like TxAgent [3] and Virtual Lab [1] that use hundreds of tools or multi-agent systems to carry out research tasks. By providing MolPerceptor as a tool to these AI scientist systems, I allow them to better understand the protein and chemical structures that they encounter in their tool outputs and reasoning processes. To evaluate these MolPerceptor-augmented AI scientists, I will test them on a set of tasks that require reasoning about chemical and molecular data, such as modifying a molecule's structure so that it is more synthesizable or binds more strongly to a protein target.

**Intellectual Merit:**

This project proposes a comprehensive, flexible framework for improving the multimodal perception abilities of foundation models. By training an AI model that can reason across diverse structural, geometric, graphical, and other data modalities, the project will significantly accelerate our ability to construct AI scientist systems that can interact with the physical world, whether through manipulating chemical structures, visualization software, or laboratory instruments.

**Broader Impacts:**

By building multimodal AI scientists for chemical and molecular discovery, I hope to apply them to a number of difficult challenges, including 1) identifying proteins involved in disease pathways by modeling the structures of these proteins and how they interact in the body and 2) engineering new protein systems that can act as biosensors (*e.g.*, to detect illicit substances) or medicines (*e.g.*, to recognize and bind to cancer cells). In general, multimodal perception allows AI scientists to serve as more physically- and chemically-informed research assistants, allowing them to perform fast computational modeling of chemical systems and identifying hidden patterns in the molecular data that a human scientist might miss, thus accelerating scientific breakthroughs.

**References:** [1] Swanson, K., *et al*. (2025). The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. Nature, 646(8085), 716–723. [2] Ghareeb, A. E., *et al*. (2025). Robin: A multi-agent system for automating scientific discovery. arXiv. [3] Gao, S., *et al*. (2025). TxAgent: An AI Agent for Therapeutic Reasoning Across a Universe of Tools. arXiv. [4] Alampara, N., *et al*. (2025). Probing the limitations of multimodal language models for chemistry and materials research. Nature Comp. Sci., 5(10), 952–961. [5] Cui, J., *et al*. (2025). L^2M^3OF: A Large Language Multimodal Model for Metal-Organic Frameworks. arXiv. [6] Li, J., *et al*. (2024). ChemVLM: Exploring the Power of Multimodal Large Language Models in Chemistry Area. arXiv. [7] Leong, S. X., *et al.* (2025). MERMaid: Universal multimodal mining of chemical reactions from PDFs using vision-language models. Matter, 102331.