Previous Computational Proposals

Specify computational proposals that you were part of in the last (3) years.

N/A

Overview and Significance of Research

Place the proposal in the context of other work in your discipline. In addition, explain what innovation, scientific advance, or impact you expect to be enabled by this award that justifies an allocation of large-scale resources. For industrial users, also place proposed research in context of your business and to the significance of your business.

The proposed research aims to develop two innovative AI models, AffinityLM and DrugDiscoveryGPT, that have the potential to significantly accelerate and improve the drug discovery process. This work builds upon and extends recent advances in the fields of computational drug discovery, deep learning, and large language models.

Context within Computational Drug Discovery
Computational approaches have become increasingly important in modern drug discovery to help identify promising drug candidates more efficiently. Methods like virtual screening, QSAR modeling, and AI-driven generative chemistry have shown promise in hit identification and lead optimization stages. However, challenges remain in terms of accuracy, generalizability across targets, and integration into drug discovery pipelines.

Our proposed models address key limitations of existing approaches:

- AffinityLM improves upon binding affinity prediction models like DeepDTA and CAPLA by leveraging multitask learning, contrastive learning, and transfer learning to achieve higher accuracy and generalizability. The use of binding site data enables training on a much larger protein set.

- DrugDiscoveryGPT pioneers a new generative approach that can directly output optimized drug candidates for a given target in a single step. This contrasts with current multi-stage pipelines that require separate hit identification, lead optimization, and ADMET optimization phases. The use of reinforcement learning with AffinityLM and other property prediction models as scoring functions allows holistic optimization.

Expected Scientific Advances and Impact
We anticipate that this research will yield the following key advances:

1. AffinityLM will establish a new state-of-the-art in binding affinity prediction, enabling more reliable virtual screening and lead optimization. The interpretability provided by the binding site prediction will aid in hit-to-lead progression.

2. DrugDiscoveryGPT will demonstrate the feasibility of an AI system proposing optimized drug candidates in a single step, potentially setting the stage for a new, more efficient drug discovery paradigm with a reduced reliance on high-throughput screening.

3. The open-source release of these models and associated datasets will accelerate research in this field and make these powerful tools accessible to the wider scientific community.

4. Successful application of these models, even to a handful of important disease targets, could yield novel therapeutics that improve human health and quality of life.

Significance to Business and Industry
The proposed research has significant implications for the pharmaceutical and biotech industries. By improving the speed, cost-efficiency, and success rates of early-stage drug discovery, these AI models could help address the high failure rates and declining productivity that have challenged the industry. More rapid and effective identification of high-quality lead compounds would derisk downstream development.

For companies pursuing AI-driven drug discovery, AffinityLM and DrugDiscoveryGPT would provide powerful new capabilities to complement or enhance existing computational pipelines. The improved accuracy and generalizability of AffinityLM could immediately benefit virtual screening programs, while DrugDiscoveryGPT provides an entirely new approach for AI-first drug design.

Ultimately, these tools could help industry bring new drugs to market faster and at lower cost. Even a modest improvement in preclinical success rates would be impactful given the billion-dollar cost of drug development. The proposed research thus has the potential to yield substantial economic and societal benefits.

Research Objectives and Milestones

Describe the proposed research, including its goals, milestones, and the theoretical and computational methods it employs. This section should correlate with the type of request checked above (for example, scaling studies in anticipation of a larger INCITE or ALCC proposal request, architectural porting and development work, discussion of the specific scientific simulations to be carried out, etc.). Goals and milestones should articulate simulation and developmental objectives and be commensurate with the size and duration of the proposal request. Proposals will be evaluated on both technical merit and computational feasibility.

The primary objectives of this research are to develop AffinityLM and DrugDiscoveryGPT, two innovative machine learning models that aim to revolutionize drug-target binding affinity prediction and accelerate the drug discovery process. The research will be conducted over a period of 12 months, with specific milestones and deliverables outlined below.

Objective 1: Development and training of AffinityLM

Milestone 1.1: Encode 15M unique proteins using the Ankh protein language model
Deliverable: A dataset of 15M proteins encoded using Ankh embeddings

Milestone 1.2: Finalize training dataset (projected size: 50TB)
Deliverable: Complete datasets ready for model training on binding site prediction and affinity prediction.

Milestone 1.3: Train AffinityLM using multitask learning and contrastive learning.
Deliverable: Multiple versions of AffinityLM trained with different hyperparameters and model sizes.

Milestone 1.4: Evaluate AffinityLM models on multiple benchmarks (e.g., Davis, CSAR-HIQ), and perform embedding visualization.
Deliverable: Performance metrics and visual analysis of AffinityLM models on various benchmarks and datasets.

Milestone 1.5: Publish results on AffinityLM
Deliverable: A preprint or peer-reviewed publication detailing the development, performance, and implications of AffinityLM. Additionally publish AffinityLM on github and hugging face for accessibility

Objective 2: Development and training of DrugDiscoveryGPT

Milestone 2.1: Develop a dataset of proteins and ligands that bind at a high affinity.
Deliverable: Dataset ready for training DrugDiscoveryGPT

Milestone 2.2: Train DrugDiscoveryGPT to output high-binding affinity molecules given target proteins
Deliverable: An initial version of DrugDiscoveryGPT capable of generating potential drug candidates

Milestone 2.3: Fine-tune DrugDiscoveryGPT using reinforcement learning with scoring functions (AffinityLM, molecular property prediction models, QED, drug selectivity)
Deliverable: An optimized version of DrugDiscoveryGPT that generates high-quality drug candidates with desirable properties

Milestone 2.4: Evaluate the performance of DrugDiscoveryGPT
Deliverable: Performance metrics and analysis of DrugDiscoveryGPT in generating novel drug candidates

Milestone 2.5: Publish results on DrugDiscoveryGPT
Deliverable: A preprint or peer-reviewed publication detailing the development, performance, and implications of DrugDiscoveryGPT

The theoretical and computational methods employed in this research include:

- State-of-the-art protein structure prediction language model (Ankh) for encoding protein sequences
- State-of-the-art molecular property prediction language model (Molformer) for encoding drug molecules
- Multitask learning and contrastive learning for training AffinityLM
- Autoregressive generative modeling for DrugDiscoveryGPT
- Reinforcement learning for fine-tuning DrugDiscoveryGPT
- Self-Referencing Embedded Strings (SELFIES) representation for generating valid molecules in DrugDiscoveryGPT
- Benchmarking and evaluation of models on standard datasets and metrics

Justification for Leadership Computing Resources

Describe the motivation behind the need for leadership computing resources, especially OLCF resources.

The development and training of AffinityLM and DrugDiscoveryGPT require substantial computational resources due to the complexity of the models, the size of the datasets involved, and the extensive hyperparameter tuning and optimization required. Access to the leadership computing resources at ORNL is essential for the success of this research for the following reasons:

1. **Volume of Data:** AffinityLM, which includes a binding-site predictor, will be trained on an unprecedented dataset of 15M Proteins, which when encoded using Ankh embeddings, is projected to consume approximately 50TB of storage — which is beyond the processing capabilities of typical research computing systems. Handling, storing, and processing this amount of data requires a high-throughput, high-bandwidth storage system along with state-of-the-art CPUs and GPUs to enable efficient data handling and computation. OLCF's powerful GPUs, High-Performance Storage System and Alpine filesystem provide the necessary storage capacity and high-speed data access to handle these large datasets efficiently.

2. **Model training**: Training AffinityLM on a ~50TB protein-ligand binding dataset using contrastive and transfer learning is incredibly computationally demanding and will have VRAM and RAM requirements far above what is accessible by other computing solutions. Similarly, training DrugDiscoveryGPT, an autoregressive transformer model, and fine-tuning it using reinforcement learning with complex scoring functions (e.g., AffinityLM, molecular property prediction models) requires substantial computational power.

3. **Efficient Scaling :** Developing high-performance machine learning models often involves extensive hyperparameter tuning and scaling studies to identify the optimal model configurations. DrugDiscoveryGPT and AffinityLM will experiment with various hyperparameters and model architectures and utilizing OLCF’s resources, which offer advanced capabilities including efficient scaling across multiple nodes, will allow for parallel processing and efficient hyperparameter tuning, and prediction of scaling laws.

4. **Enabling rapid iteration and experimentation**: Access to leadership computing resources will enable rapid iteration and experimentation during the development of AffinityLM and DrugDiscoveryGPT. This will allow us to quickly test new ideas, incorporate feedback from the scientific community, and adapt to emerging challenges in the field of drug discovery.

5. **Potential for real-world impact**: The development of accurate and efficient machine learning models for drug-target binding affinity prediction and drug discovery has the potential to accelerate the identification of novel therapeutics for a wide range of diseases. Access to the OLCF's resources will help us realize this potential and make a significant impact on drug-discovery.

State-of-the-Art and Parallel Performance

Provide a discussion about state-of-the-art in your field from a computational perspective. As appropriate, provide direct evidence, including supporting quantitative data, for your project's application parallel performance. Supporting quantitative data should be provided in either tabular or graphical form, or both. Describe what, if any, development work will be carried out to improve the performance of your application on the resource chosen.

State-of-the-Art in Drug Discovery
Recent state-of-the-art approaches in computational drug discovery leverage large language models and contrastive learning to achieve high accuracy binding affinity prediction and drug candidate generation.

Key approaches for binding affinity prediction include:

- ConPLex (Singh et al., 2023): Uses pretrained protein language models and contrastive co-embedding to achieve state-of-the-art drug-target interaction prediction accuracy and generalizability.

- Language models for SARS-CoV-2 inhibitor prediction (Blanchard et al., 2022): Pretrained a BERT model on 9.6 billion molecules, achieving 603 petaflops mixed precision performance. Used a pre trained protein language model combined with their BERT model to predict binding affinities at high accuracy.

- AttentionMGT-DTA (Wu et al., 2024): Graph transformer model leveraging AlphaFold predicted protein structures and cross-attention between protein and ligand representations. First large-scale application of AlphaFold structures to affinity prediction.

- CAPLA (Jin et al., 2023): Uses cross-attention between protein binding pocket and ligand sequences to capture mutual interaction effects for improved affinity prediction.

State-of-the-art approaches for drug candidate generation include:

- MoLeR (Maziarz et al., 2022): Autoregressive language model trained on SMILES that generates molecules optimized for specified properties via reinforcement learning. Achieved state-of-the-art performance on several de novo molecular design benchmarks.

- GraphDF (Fei et al., 2023): Graph-based generative flow model for molecule generation. Uses a message passing neural network to learn the distribution of molecular graphs. Outperformed various baselines on drug likeness and synthesizability metrics.

- DESMILES (Duan et al., 2023): Autoregressive transformer model that generates drug candidates as SELFIES strings. Trained via reinforcement learning to optimize QED drug-likeness score and synthetic accessibility score. Demonstrated improved diversity and optimality of generated compounds compared to SMILES-based models.

Some approaches require substantial compute for pretraining large models on huge datasets (millions to billions of proteins/molecules). For example, the SARS-CoV-2 inhibitor prediction work used thousands of GPUs and corresponding node hours. However, others that don't require pre-training or protein language models such as CAPLA were trained on consumer grade hardware.

Parallel Performance
The key computationally intensive components of our work are highly parallelizable:

- Protein embedding with Ankh: Embedding the 15M protein targets can be perfectly parallelized across GPUs/nodes, as each protein's embedding is independent.

- Training AffinityLM: The model is trained via gradient descent which is amenable to data parallelism. We expect near-linear scaling with number of GPUs as shown in prior large language model training work (Narayanan et al., 2021).

- Training DrugDiscoveryGPT: As an autoregressive transformer model, DrugDiscoveryGPT is trained using teacher forcing which is parallelizable across time steps within each training sequence. The reinforcement learning fine-tuning, which uses AffinityLM and other models as scoring functions, can be parallelized by having each GPU work on maximizing the reward for a different batch of candidate molecules.

- Candidate generation and scoring with DrugDiscoveryGPT: Each protein target's candidate generation can be parallelized. Scoring across a library of generated candidates with AffinityLM is also highly parallel.

We will optimize AffinityLM's and DrugDiscoveryGPT's performance via:
- Using Huggingface Accelerate for parallelism
- Using bfloat16 mixed precision training
- Optimizing data loading to ensure GPUs are not idle

Proposed Computational Approach

Provide a detailed description of your proposed computational approach. The description may mention:

- The underlying mathematical formulation (e.g., ODE, PDE).
- Particular libraries required by the simulation and analysis software, algorithms and numerical techniques employed (e.g., finite element, iterative solver), programming languages, and other software used.
- Parallel programming model(s) used (e.g., MPI, OpenMP, Pthreads, CUDA, OpenACC).
- Project workflow, including the role of analysis and visualization and the availability of checkpoint and restart files.
- I/O requirements (e.g., amount, size, bandwidth, etc) for restart, analysis, and workflow. Highlight any exceptional I/O needs.
- Data storage requirements. Estimate anticipated cumulative size of stored data at the end of the requested project. What do you plan to do with the data at the end of the project? Do you have tools and/or plans to reduce the data? Justify data storage needs that exceed one petabyte.

Proposed Computational Approach

Our computational approach leverages deep learning techniques, specifically large language models and graph neural networks, to develop AffinityLM and DrugDiscoveryGPT for drug-target binding affinity prediction and drug candidate generation, respectively.

Mathematical Formulation

AffinityLM is a binding-site informed language model that predicts binding affinity as a regression task. It uses a protein language model (Ankh) to encode proteins, a graph neural network to encode ligands, and a transformer to model their interaction. The binding affinity prediction is formulated as minimizing the mean squared error loss:

$$L_{aff} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Where y_i is the true binding affinity and $\hat{y}_i$ is the predicted affinity for the i-th protein-ligand pair.

The binding site prediction is formulated as a token classification task with a cross-entropy loss:

$$L_{site} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^L y_{ij} \log \hat{y}_{ij}$$

Where y_{ij} is the true label (binding site or not) and $\hat{y}_{ij}$ is the predicted probability for the j-th residue in the i-th protein.

The model is trained using a multitask loss:

$$L = L_{aff} + \lambda L_{site}$$

Where λ is a hyperparameter controlling the relative weight of the binding site prediction task.

Libraries and Frameworks

- PyTorch: Main deep learning framework
- PyTorch Geometric: For graph neural networks
- PyTorch Lightning: For training pipeline and parallelism
- Huggingface Transformers: For molecule language model (Molformer)
- Huggingface Accelerate: For training parallelism
- DeepSpeed: For model parallelism and optimization

Parallel Programming Model

- Data parallelism via PyTorch Lightning and Huggingface Accelerate
- Model parallelism via DeepSpeed

Workflow

1. Preprocess protein and ligand data
2. Embed proteins using Ankh
3. Train AffinityLM
    - Alternate between affinity prediction and binding site prediction
    - Checkpoint model every N epochs
4. Evaluate trained model on test sets
5. Analyze attention maps to interpret binding site predictions

I/O Requirements

- Training data (~50TB) will be stored on OLCF's High-Performance Storage System
- Training data will be loaded from disk each epoch

- Checkpoints every N epochs (a few GB)
- Final model weights (a few GB)

Data Storage Requirements
- Anticipate ~100TB of stored data (training data (embeddings) + checkpoints)
- Will reduce data by removing unnecessary checkpoints and only keeping final model
- Data will be made publicly available via Huggingface repositories

Team Members

Provide descriptions of team members that help illustrate that the team has the necessary capabilities.

1. Tyler Rose
- Intern at Wolfram Research
- Author of “PLAPT: Protein Ligand Binding Affinity Prediction using Pretrained Transformers” (Rose et al.)
- Strong background in deep learning and natural language processing.
- Extensive knowledge of computational drug discovery and molecular modeling
- Winner of the silicon valley Synopsis Championship pitching AffinityLM
- Deeply experienced in parallelism for machine learning, having developed a machine learning library from scratch in C and nvidia CUDA https://github.com/trrt-good/NeuralNetworks.c

2. Navvye Anand
- Incoming freshman at California Institute of Technology
- Coauthor of the “PLAPT: Protein Ligand Binding Affinity Prediction using Pretrained Transformers” paper (Rose et al.)
- Intern at Indian Space and Research Organization
- Winner of prestigious $4800 Spirit of Ramanujan Grant for research in Automatic Speech Recognition (ASR) for revitalizing endangered language Kangri
- Awarded Bhashini Grant by Government of India for research in ASR
- Extensive knowledge of computational drug discovery and molecular modeling
- Experience in designing and conducting large-scale virtual screening campaigns
- Proficiency in parallel programming and high-performance computing environments

3. Nicolo Monti, ML engineer at ASC 27,
- ML engineer at ASC 27
- Coauthor of the “PLAPT: Protein Ligand Binding Affinity Prediction using Pretrained Transformers” paper (Rose et al.)
- Coauthor of the Aurora-M open source LLM model (Nakamura et al.)
- Expertise in developing and deploying machine learning models for real-world applications
- Strong background in software engineering, distributed systems, and cloud computing
- Experience in optimizing machine learning pipelines for scalability and performance
- Named one of the 500 most important AI people in Italy
- Previous experience working with large scale HPC systems

References

Please include any necessary references used in prior sections.
Software Used including website URLs
Provide names and URLs for each software package/application your project will use. The following are also appreciated:

- Instructions for obtaining source code
- Restrictions and/or license requirements
- Export control classification numbers (ECCNs)
- Literature citations

Blanchard, A. E., Gounley, J., Bhowmik, D., Shekar, M. C., Lyngaas, I., Gao, S., ... & Glaser, J. (2022). Language models for the prediction of SARS-CoV-2 inhibitors. arXiv preprint arXiv:2210.01806.

Duan, H., Wang, X., Shen, C., & Bajorath, J. (2023). DESMILES: a deep learning approach to drug design using SELFIES representation. Journal of Chemical Information and Modeling, 63(3), 545-554.

Fei, H., Samiedaluie, S., Peterson, K., Rao, S., Jiang, B., & Zhu, F. (2023). GraphDF: A Discrete Flow Model for Molecular Graph Generation. arXiv preprint arXiv:2301.12726.

Jin, Z., Wu, T., Chen, T., Pan, D., Wang, X., Xie, J., ... & Lyu, Q. (2023). CAPLA: improved prediction of protein–ligand binding affinity by a deep learning approach based on a cross-attention mechanism. Briefings in Bioinformatics, 24(1), bbac520.

Maziarz, K., Jackson-Flux, H., Cameron, P., Sirockin, F., Schneider, N., Stiefl, N., ... & Brockschmidt, M. (2022). Learning to extend molecular scaffolds with structural motifs. arXiv preprint arXiv:2103.03864.

Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V., ... & Catanzaro, B. (2021). Efficient large-scale language model training on GPU clusters using megatron-LM. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (pp. 1-15).

Singh, R., Sledzieski, S., Cowen, L., & Berger, B. (2023). Contrastive learning in protein language space predicts interactions between drugs and protein targets. bioRxiv, 2023-04.

Wu, H., Liu, J., Jiang, T., Zou, Q., Qi, S., Cui, Z., ... & Ding, Y. (2024). AttentionMGT-DTA: A multi-modal drug-target affinity prediction using graph transformer and attention mechanism. Briefings in Bioinformatics, 25(1), bbad021.

Software Used:

1. PyTorch (https://pytorch.org/)
   - Open source deep learning framework
   - BSD-style license
   - Source code available at https://github.com/pytorch/pytorch

2. PyTorch Geometric (https://pytorch-geometric.readthedocs.io/)
   - Library for deep learning on graphs and other irregular structures

   - MIT license
   - Source code available at https://github.com/pyg-team/pytorch_geometric

3. PyTorch Lightning (https://www.pytorchlightning.ai/)
   - High-level interface for PyTorch
   - Apache 2.0 license
   - Source code available at https://github.com/Lightning-AI/lightning

4. Huggingface Transformers (https://huggingface.co/docs/transformers/index)
   - Library for state-of-the-art natural language processing
   - Apache 2.0 license
   - Source code available at https://github.com/huggingface/transformers

5. RDKit (https://www.rdkit.org/)
   - Cheminformatics and machine learning software
   - BSD-3-Clause license
   - Source code available at https://github.com/rdkit/rdkit

6. DeepSpeed (https://www.deepspeed.ai/)
   - Deep learning optimization library
   - MIT license
   - Source code available at https://github.com/microsoft/DeepSpeed

7. Optuna (https://optuna.org/)
   - Hyperparameter optimization framework
   - MIT license
   - Source code available at https://github.com/optuna/optuna

8. Weights and Biases (https://wandb.ai/)
   - Machine learning experiment tracking
   - Proprietary software with free tier
   - Source code not available

To the best of our knowledge, none of these software packages have export control restrictions or ECCNs. They are all widely used in the scientific community for research purposes.