**Topic of the Task:**

Develop a novel algorithm based on deep learning for AI-Powered Drug Discovery and Development. Implement cross-validation using other datasets to ensure robustness and accuracy.

**Research Report:**

1. ***Traversing Chemical Space with Active Deep Learning for Low-Data Drug Discovery***

* **Link:** <https://www.nature.com/articles/s43588-024-00697-2>
* **Dataset:** Three large-scale molecular libraries for virtual screening in low-data settings.
* **Methodology or Models Used:** Deep neural networks with six active learning strategies to iteratively select molecules for labelling.
* **Novelty:** Demonstrated up to 6× improvement in hit discovery rates compared to brute-force screening in low-data contexts.
* **Accuracy (%):** Relative hit rate improvement (~600% increase); not expressed as classification accuracy.
* **Evaluation metrics:** Hit discovery rate, screening efficiency.

1. ***PILOT: Equivariant Diffusion for Pocket‑Conditioned de novo Ligand Generation***

* **Link:** <https://pubs.rsc.org/en/content/articlelanding/2024/sc/d4sc03523b>
* **Dataset:** CrossDocked2020 benchmark; Kinodata‑3D for unseen kinase pockets.
* **Methodology or Models Used:** Equivariant diffusion model conditioned on protein pockets, with importance sampling for multi-objective guidance.
* **Novelty:** First pocket-conditioned equivariant diffusion model using importance sampling to balance binding affinity & synthetic accessibility.
* **Accuracy (%):** Demonstrated significant outperformance over prior methods on CrossDocked2020.
* **Evaluation metrics:** RMSD, predicted IC₅₀ values, binding affinity ranking accuracy.

1. ***Learning to Navigate the Synthetically Accessible Chemical Space Using Reinforcement Learning***

* **Link:** <https://arxiv.org/abs/2004.12485>
* **Dataset:** Commercially available building-block molecules, tested on three HIV targets.
* **Methodology or Models Used:** Policy-gradient reinforcement learning for multi-step virtual synthesis (PGFS).
* **Novelty:** Directly incorporates synthetic accessibility in molecule generation.
* **Accuracy (%):** Achieved top-tier QED and penalized clogP; validated with HIV-target proxies.
* **Evaluation metrics:** QED (quantitative drug-likeness), penalized clogP, target activity score.

1. ***Automatic Chemical Design Using a Data‑Driven Continuous Representation of Molecules***

* **Link:** <https://arxiv.org/abs/1610.02415>
* **Dataset:** Hundreds of thousands of small, drug-like molecules (<9 heavy atoms).
* **Methodology or Models Used:** Variational autoencoder with an encoder–decoder–predictor architecture.
* **Novelty:** Enables gradient-based optimization in a continuous latent space.
* **Accuracy (%):** Demonstrated generation of valid novel molecules and property optimization; exact percent not provided.
* **Evaluation metrics:** Validity, novelty, property improvement metrics.

1. ***A Deep‑Learning View of Chemical Space Designed to Facilitate Drug Discovery (DESMILES)***

* **Link:** <https://arxiv.org/abs/2002.02948>
* **Dataset:** Benchmark sets for D2 receptor inhibition and additional receptor docking cases.
* **Methodology or Models Used:** Deep neural network (DESMILES) for iterative molecule property refinement.
* **Novelty:** Reduced failure rate by 77% versus previous models in D2 receptor tasks.
* **Accuracy (%):** 77% decrease in failure rate.
* **Evaluation metrics:** Failure rate, docking binding scores.

1. ***Screening ultra‑large virtual libraries using the V‑SYNTHES Workflow.***

* **Link:** <https://www.nature.com/articles/d41573-022-00002-8>
* **Dataset:** Enamine REAL Space library (~hundreds of millions of compounds) screened against CB1 and CB2 receptors.
* **Methodology or Models Used:** Two-stage docking — fragment-based initial screening followed by full-compound docking optimized for scale.
* **Novelty:** Enabled ultra-large library screening with ~5,000× speed-up; achieved **33% functional hit rate** (Ki < 10 µM) compared to ~17% in conventional screens.
* **Accuracy (%):** 33% experimental hit rate (Ki < 10 µM).
* **Evaluation metrics:** Hit rate, docking score enrichment, scaffold diversity.

1. ***VirtualFlow: Open‑Source Platform for Ultra‑Large Docking***

* **Link:** <https://www.nature.com/articles/s41592-021-01229-5>
* **Dataset:** 1.3 billion compounds screened using QuickVina 2.
* **Methodology or Models Used:** Containerized, massively parallel docking workflow for scalable virtual screening on HPC and cloud platforms.
* **Novelty:** Open-source framework enabling accessible ultra-large docking campaigns.
* **Accuracy (%):** Active hits experimentally confirmed for Keap1 protein; specific percentage not provided.
* **Evaluation metrics:** Docking score rankings, experimental binding confirmation.

1. ***Artificial Intelligence–Enabled Virtual Screening of Ultra‑Large Chemical Libraries with Deep Docking***

* **Link:** <https://www.nature.com/articles/s41596-021-00577-0>
* **Dataset:** ZINC20 database containing billions of small molecules.
* **Methodology or Models Used:** Iterative surrogate machine learning models combined with docking in an active learning loop.
* **Novelty:** Achieved ~100× screening speed-up and 1,000× hit enrichment compared to brute-force virtual screening.
* **Accuracy (%):** Hundreds- to thousands-fold hit enrichment.
* **Evaluation metrics:** Hit enrichment factor, screening acceleration.

1. ***Large‑Scale Pretraining Improves Sample Efficiency of Active Learning‑Based Molecule Virtual Screening***

* **Link:** <https://arxiv.org/abs/2309.11687>
* **Dataset:** Subset of ~99.5 million molecules from ultra-large chemical libraries.
* **Methodology or Models Used:** Pretrained transformers and graph neural networks used as surrogate models in a Bayesian active learning pipeline.
* **Novelty:** Recovered **58.97%** of the top 50,000 docking hits after screening only **0.6%** of the dataset — an 8% improvement over prior approaches.
* **Accuracy (%):** 58.97% recall of top-ranked hits.
* **Evaluation metrics:** Recall of top candidates, fraction of library screened.

1. ***A Novel Framework Integrating AI Model and Enzymological Experiments Promotes Identification of SARS‑CoV‑2 3CL Protease Inhibitors and Activity‑Based Probe***

* **Link:** <https://arxiv.org/abs/2105.14224>
* **Dataset:** Bioactive chemical library containing commercially available small molecules screened against SARS‑CoV‑2 3CL protease.
* **Methodology or Models Used:** AI-prediction model integrated with in-silico docking, followed by enzymological validation experiments (iterative rounds).
* **Novelty:** Combines AI-driven virtual screening with enzymatic assays; identified six novel 3CL protease inhibitors, four with IC₅₀ < 3 µM; also discovered an activity-based probe.
* **Accuracy (%):** 29.41% experimental hit rate in enzymatic assays.
* **Evaluation metrics:** Hit rate, IC₅₀ values, activity-based probe identification.