Sigma-Aldrich®
Lab & Production Materials

AIDDISON™: Using AI-Powered Software to Accelerate and Optimize Novel Drug Design

INTRODUCTION

The popularity and widespread use of artificial intelligence (AI)-based methods has increased dramatically over the last decade and has touched almost every facet of life sciences, including drug discovery. AI application in the chemistry domain has seen an exponential increase in research publications since 2015 (Baum, 2021). Some of the most promising areas deal with the prediction of the bioactivity of novel molecules, 3D protein structures from sequence data, calculation of a variety of ADME-Tox properties, and suggestions of synthetic routes to complex target molecules. Advances in virtual screening have made efficient sampling of vast chemical spaces (trillions of molecules) possible, thus providing a rich source of novel and, equally important, synthesizable molecules for evaluation in machine learning (ML) models or docking experiments (Howes, 2022). However, most of these tools are not readily available to bench scientists using a simple, consistent, and easily accessible interface.

AIDDISON™, an AI-powered drug discovery platform, was developed to address this need and accelerate the hit finding and optimization process in drug discovery. AIDDISON™ harnesses the power of AI and computational tools to virtually screen for novel molecules with machine learning (ML) models to predict pharmacokinetic profiles and design *de novo* molecules with AI generative methods. AIDDISON™ uses a scientific approach to first generate thousands of viable candidate molecules as the starting point in an analysis. These can come from a variety of sources – 2D similarity and 2D pharmacophore searches, de novo design using generative models and, of course, direct user input. They can be filtered by computed molecular properties to identify those with optimal properties. Filtering by ADME-Tox properties earlier in the research pipeline has been shown to lead to fewer failures later due to a poor pharmacokinetic (PK) profile (Daina, 2017). Shape-based alignment versus a known active ligand and docking experiments are used to evaluate potential biological activity. Finally, the best molecular designs could be sent to SYNTHIA™ retrosynthesis software to assess their synthesizability and identify necessary reagents.

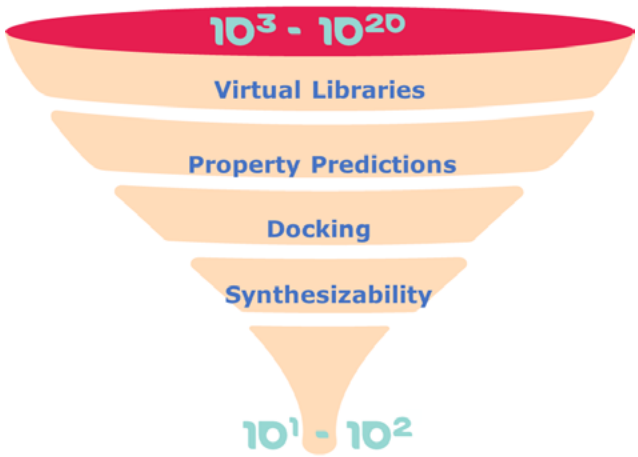

Figure 1. Computer-aided drug discovery funnel from virtual chemical space to small number of compounds to synthesize.

AIDDISON™ has a secure software as a service (SaaS) environment and cloud-based serverless architecture, which allows companies to keep computational costs under control while giving access to proven, commercial AI and computational technologies to end users. To assist bench scientists, the interface is simple and clean with complex tasks automated, workflows streamlined, and machine learning models readily incorporated. The results are then visually displayed in meaningful and intuitive ways to enable scientists to innovate and unlock their own ingenuity.

FEATURES

Projects

Projects in AIDDISON™ are high level “folder” objects that include information about the project itself, molecules, their computed properties, and results from a variety of workflows and filters. They are quite useful for organizing ideas around a common theme, task or set of molecules. Projects can be added and deleted at will. They can also be shared with other members of your organization to facilitate collaboration.

Molecule Library

In each project, AIDDISON™ uses molecular collections to handle groups of molecules within the system. Each collection subtype has its own unique use and features. For example, the Molecule Library is the only non-removeable set present per project. It is a special case of the molecule collection object since all other behaviors are the same. Molecules can be added here, and a variety of workflows can be initiated from any individual molecule. Basically, this is the “home page” for a project as users can always go back there as a reference point before switching to other features. Molecular structures can be added via a structure editor, pasting a SMILES string, or from a CSV or SD-File. They can be exported to a CSV or SD-File or copied into the Molecule Library or a Molecule Set for further processing.

Molecule Sets

Molecule Sets are a convenient way of assembling data sets from a variety of sources. They can come from the Molecule Library, Job Results, Filtered Results, or even other Molecule Sets. This editable list of molecules with configurable properties (using Settings) can be shared with other projects or discarded if need be.

Filters, Settings, and Filtered Sets

Molecular property filters are routinely used to reduce the number of molecules sent as input to the more computationally expensive algorithms – 3D Shape-based Search and Molecular Docking. Filters can only be applied to the Molecule Library and Molecule Sets. Defined substructures can be used to include or exclude molecules. Settings are used to add molecular properties such as results from machine learning (ML) models or multi-parameter optimization (MPO) calculations to the table for comparison. Support is readily available to assist the user in interpreting the calculations. Ultimately, the filtered results are stored in a molecule collection called Filtered Sets, which contains information about the source of the molecules in the set and the filters used to generate it.

ML Models

The lack of quality machine learning ADME-Tox models is a significant bottleneck in the lead optimization process. This is usually due to a lack of reliable data. However, while large, *in vivo* data sets have recently become available, they are typically confined to large pharma companies only. ADME-Tox ML models are built by Merck KGaA, using proprietary, in-house data. Though the data behind these models is unavailable for public consumption, the ML models themselves have been incorporated in AIDDISON™ and can be used to optimize the choice of molecules for further consideration. Currently, they include aqueous solubility, CaCo2 permeability, hERG activity, HepG2 cytotoxicity, intrinsic clearance in a human, rat, or mouse, as well as an estimation of the fraction unbound in each species.

Figure 2. Table view of a molecule set with computed properties color-coded by ADMET warning levels.

Glowing Twin

In addition to the ADME-Tox ML models provided, Glowing Twins give a more chemically intuitive view of the results. Once a particular ML is chosen, the atoms in the molecule that significantly affect the property calculation (good or bad) are displayed. The structure then can be manipulated using the provided editor and a new Glowing Twin molecule is displayed showing how the changes made to the original molecule affect the calculated property.

IDEA GENERATORS

Similarity Search

All AIDDISON™ workflows are seamlessly integrated. The first and most recognizable method, Similarity Search, uses FPSim2 to rapidly perform chemical similarity search of a given target molecule on large collections. Structural features (ECFP2) are encoded into molecular fingerprints (bitstrings) for comparison. Tanimoto similarity is then computed to target structure. Those molecules whose similarity is above the set threshold are returned. Currently, the available collections include Sigma-Aldrich® Catalog, PubChem, ChemBL and Zinc.

User-defined collections can be added as well. The Job Results for a Similarity Search are reported in a Table View format where Similarity is the first column. Neighbor Search employs this same technology to search for similar molecules within a given project.

Virtual Chemical Spaces

During the last few years, the number of available screening compounds available has grown larger than ever before, through both physical and virtual libraries (Howes, 2022). Billions of synthetically available compounds are offered by companies, like Enamine and WuXi. Even larger are the in-house virtual libraries, or spaces, owned by large pharmaceutical companies.

For example, Merck KGaA has built the Merck Accessible Inventory (MASSIV), a virtual space of ~10^{20} molecules. AstraZeneca built a virtual chemical space of ~10^{17} possible compounds using Enamine building blocks (Hoffman, 2019).

SA-Space™ is a synthetically accessible, ultra-large virtual chemical space built on Sigma-Aldrich® building block chemicals and well-known robust chemical transformation rules. SA-Space™ encompasses approximately 25 billion virtual compounds which can be searched quickly through AIDDISON™, AI-powered drug discovery SaaS platform. For each search in SA-Space™, users are provided with links to all purchasable reagents for synthesis and synthesis protocol details to execute in the labs. AIDDISON™ seamlessly integrates the computational tools, access to vast chemical space, and direct connection to synthetic planning and chemical catalogs to enable scientists to achieve their goals.

Pharmacophore Search

Pharmacophore Search is a 2D similarity method where similarity to a target molecule is computed by comparing Feature Trees (FTrees), as implemented by BioSolveIT (BioSolveIT, 2022). This is a highly efficient and effective tool for scaffold hopping and ligand-based screening of incredibly vast virtual chemical space. Its underlying topological descriptors capture ring/chain and pharmacophore attributes and by using a reduced graph representation, the relationships amongst them are kept intact allowing for extremely rapid comparisons. Pharmacophore Search results are reported in the Gallery View format where molecules can be visually compared to the target and structural features directly matched. Alternatively, results can be displayed in Table View which allows the user to manipulate the order of molecules or save to a molecule set for further refinements. SA-Space™ (by Sigma-Aldrich® and only available in AIDDISON™), Real (by Enamine) and GalaXi (by WuXi) are available to search, yielding billions of possibilities!

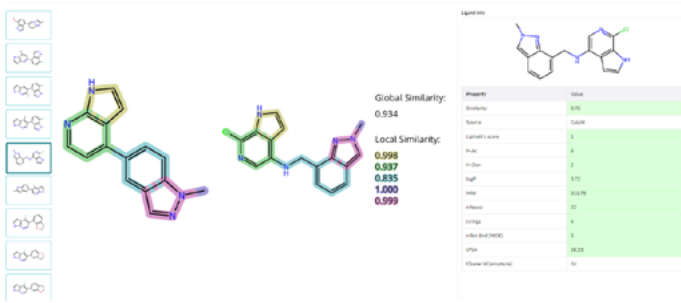

Figure 3. Gallery View of Pharmacophore (FTrees) Search results. Colors identify similar functional groups and the degree of similarity between target (left) and

De novo Design (Generative Models)

De novo design of molecules employing generative AI models has become increasingly popular due to their use in drug discovery applications. Within AIDDISON™, *de novo* design, based upon REINVENT (Blaschke, 2020), is used to generate a set of virtual molecules with desired chemical properties from a target molecule of interest. Generated molecules are scored based on novelty, drug-likeness, similarity to target structure (using FTrees) and an assessment of synthetic accessibility via a connection to SYNTHIA™ retrosynthesis software. This is a feature unique to AIDDISON™. Several options are available to generate different types of molecules that are saved and available for further refinements.

Clustering

In addition to filtering steps, clustering or diversity sampling is another effective way of reducing large sets of molecules to a more manageable number for evaluation by identifying representative subsets. There are several options available in AIDDISON™. Traditional hierarchical and Jarvis-Patrick (Butina) clustering as well as MinMax and Sphere Exclusion sampling have been implemented. Clustering molecules using molecular properties is also provided to identify novel molecules with similar properties but possibly different topologies. Dimension Reduction, based on a user-define MPO, is used to accomplish this task. The Grid View is used to visualize the results via structure cards of centroids and clusters.

EVALUATORS

Shape-based Search

Shape-based Search, based on Cresset's Flare (Cheeseright, 2022), is an effective method for evaluating novel molecules based upon a 3D alignment to a target molecule. Once the low-energy conformer has been determined, property-based field points are generated for it.

These field point patterns, combined with their shape, are used to align and score a 'database' of molecules against a reference that is usually a known active. In this context, the default Dice similarity has worked well and provided meaningful results. The Gallery View is then used to visualize the 3D structural alignment results. Alternatively, using the asymmetric Tversky index, a user can specify whether a sub-field (akin to a substructure) or super-field search is performed. For example, these can be useful in determining whether a novel bio-isosteric fragment is found in a larger target molecule. Another practical use is to reduce the number of molecules sent to molecular docking by *a priori* filtering out molecules that don't align well to a reference ligand.

Molecular Docking

Protein structures are necessary to initiate the Molecular Docking workflow. They can be easily added to a company's protein collection by selecting Add Protein in the docking workflow or by accessing the protein collection directly. Protein Data Bank (PDB) accession codes can be entered and AIDDISON™ retrieves the requested protein directly from the PDB. Alternatively, a user can directly upload a protein structure (in PDB format) from their own system. Ligands and the active site about them within 6Å are automatically identified. Possible alternative binding pockets, which most likely do not have a bound ligand, can be algorithmically identified as well.

Accurate modeling of protein-ligand interactions is essential for successful docking experiments. The Molecular Docking workflow, Flare Docking from Cresset (Flare Docking, 2022), is used to evaluate molecules in an active site or alternative binding pocket of a protein. Upon selection of the target protein structure, hydrogens and charges are added to amino acid residues automatically (via Cresset tools). The binding pocket (active site in most cases) is selected by either choosing a reference ligand provided in the PDB file or an unoccupied pocket identified algorithmically. Next, a user-defined set of molecules is designated for docking. Hydrogens and appropriate charges are set for each molecule and multiple low-energy conformers are generated for these molecules prior to docking.

Results can be visualized in the Gallery View, where the new molecules are displayed inside the chosen pocket. The solvent-accessible surface of the protein and the reference ligand can be displayed as well, which quickly allows for the ability to see how each molecule fits in the pocket relative to the crystal structure. The Table View gives the user the chance to sort by various computed quantities, such as VSScore (virtual screen score), which provides an indication of potential bioactivity; or RankScore, which specifies which molecule (via its docked pose) is most like the ligand structure, and a calculated Free Energy of Binding (dG). Resulting 3D structures can be saved as an SD-File for future use.

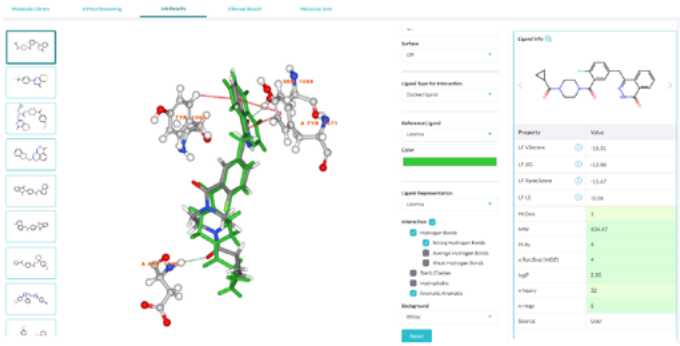

Figure 4. Gallery View of Molecular Docking results. A molecule is docked in the active site of the target protein and the literature reference ligand (green) is also shown. The interactions between binding site residues and the docked and/or reference ligand(s) can be visualized to evaluate the binding pose of the docked molecule.

SYNTHIA™ and Shopping Cart

The capability of converting those interesting but still theoretical molecules into actual compounds is another advantage provided by AIDDISON™. This is accomplished by a direct connection to SYNTHIA™ retrosynthesis software. The best possible synthetic route for execution, *i.e.*, one with minimal costs and number of steps but still has the best chance of success, can dramatically reduce the time it takes the chemist to go from ideation to a physical sample for testing. This, coupled with easy access to the Sigma-Aldrich® catalog via an online Shopping Cart, increases the likelihood that molecules designed by AI and other computational tools can be turned into real compounds in a short period of time.

SECURITY

AIDDISON™ is a zero-footprint cloud application and ISO 27001 certified for information security management systems (ISMS). All data associated with any project is protected and only accessible by members of the same company. This state-of-the-art secure and scalable cloud infrastructure allows cost-efficient and seamless access through a convenient user-interface.

SUMMARY

AIDDISON™ combines AI/ML methods and computer-aided drug design (CADD) tools, both developed in-house and licensed from well-known commercial partners, into a valuable toolkit for virtual screening, scaffold-hopping, and lead identification in medicinal chemistry. AIDDISON™ incorporates SA-spaceTM, a synthetically accessible chemical space of approximately 25 billion virtual compounds built on the Sigma-Aldrich® catalog of molecules, that are readily available for purchase, and well-known robust chemical transformation rules. The platform also integrates SYNTHIA™ retrosynthesis software to assess the synthesizability of the best molecular designs.

Learn more about AIDDISON™

Unlock Your Ingenuity!

AUTHORS

Ashwini Ghogare, Ph.D., is a leader in transformational drug discovery technologies with extensive experience in software development, new business incubation, and software product sales. She is the Head of AI and Automation in Drug Discovery at Merck KGaA, Darmstadt, Germany.

Churl Oh, Ph.D., is an expert commercial scientific software researcher and developer specialized in chemistry and cheminformatics solutions for over two decades. He is the Head of Software R&D in drug discovery at MilliporeSigma, a business of Merck KGaA, Darmstadt, Germany.

Mohammad Rezai, Ph.D. is a cheminformatics specialist and Application Scientist for AIDDISON, at Merck KGaA, Darmstadt, Germany

TRADEMARK STATEMENT

References

Baum, A. (2021). Artificial Intelligence in Chemistry: Current Trends and Future Directions. *J. Chem. Inf. Model.*, 61, 3197-3212. doi:https://doi.org/10.1021/acs.jcim.1c00619

BioSolveIT. (2022, 3). *FTrees - Similarity.* Retrieved from https://www.biosolveit.de/wp-content/uploads/2022/03/FTrees.pdf

Blaschke, T. (2020). REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inf. Model.*, 60, 5918-5922. doi:https://doi.org/10.1021/acs.jcim.0c00915

Cheeseright, T. (2022). *Fragment hopping with Blaze*. Retrieved from https://www.cresset- group.com/media/uploads/files/Fragment-hopping-with-Blaze.pdf

Daina, A. (2017). SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep, 7, 42717. doi:https://doi.org/10.1038/srep42717

Flare Docking. (2022). Retrieved from Flare Docking

Hoffman, T. (2019). The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discovery Today*, 24(5), 1148-1156. doi:https://doi.org/10.1016/j.drudis.2019.02.013

Howes, L. (2022). Hunting for drugs in chemical space. *Chemical and Engineering News*, 100(22). Retrieved from https://cen.acs.org/pharmaceuticals/drug-discovery/Hunting-drugs-chemical- space/100/i23

Merck KGaA
Frankfurter Strasse 250
64293 Darmstadt, Germany

MK_WP9256EN Ver. 2.0
08/2023