pharmaceuticals

MDPI

Article

A Deep-Learning Proteomic-Scale Approach for Drug Design

Brennan Overhoff, Zackary Falls, William Mangione and Ram Samudrala *

Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY 14203, USA; brennano@buffalo.edu (B.O.); zmfalls@buffalo.edu (Z.F.); wmangion@buffalo.edu (W.M.)

* Correspondence: ram@compbio.org

Abstract: Computational approaches have accelerated novel therapeutic discovery in recent decades. The Computational Analysis of Novel Drug Opportunities (CANDO) platform for shotgun multi-target therapeutic discovery, repurposing, and design aims to improve their efficacy and safety by employing a holistic approach that computes interaction signatures between every drug/compound and a large library of non-redundant protein structures corresponding to the human proteome fold space. These signatures are compared and analyzed to determine if a given drug/compound is efficacious and safe for a given indication/disease. In this study, we used a deep learning-based autoencoder to first reduce the dimensionality of CANDO-computed drug–proteome interaction signatures. We then employed a reduced conditional variational autoencoder to generate novel drug-like compounds when given a target encoded "objective" signature. Using this approach, we designed compounds to recreate the interaction signatures for twenty approved and experimental drugs and showed that 16/20 designed compounds were predicted to be significantly (p-value ≤ 0.05) more behaviorally similar relative to all corresponding controls, and 20/20 were predicted to be more behaviorally similar relative to a random control. We further observed that redesigns of objectives developed via rational drug design performed significantly better than those derived from natural sources (p-value ≤ 0.05), suggesting that the model learned an abstraction of rational drug design. We also show that the designed compounds are structurally diverse and synthetically feasible when compared to their respective objective drugs despite consistently high predicted behavioral similarity. Finally, we generated new designs that enhanced thirteen drugs/compounds associated with non-small cell lung cancer and anti-aging properties using their predicted proteomic interaction signatures. his study represents a significant step forward in automating holistic therapeutic design with machine learning, enabling the rapid generation of novel, effective, and safe drug leads for any indication.

Keywords: computational drug design; deep learning; multiscale; polypharmacology; autoencoder; docking; recurrent neural network

Citation: Overhoff, B.; Falls, Z.; Mangione, W.; Samudrala, R. A Deep-Learning Proteomic-Scale Approach for Drug Design. *Pharmaceuticals* **2021**, *14*, 1277. https://doi.org/10.3390/ph14121277

Academic Editor: Osvaldo Andrade Santos-Filho

Received: 24 September 2021
Accepted: 29 November 2021
Published: 7 December 2021

1. Introduction

Drug discovery—identifying chemicals with therapeutic effects against a particular indication/disease that is safe for human use—is a long, laborious, and expensive process. On average, $3 billion and about 15 years are required to bring a novel chemical entity to the market using traditional approaches [1]. Computational methods are a popular means of identifying potential leads through paradigms such as high-throughput virtual screening [2–5], where simulations are run to assess the binding affinity of a library of compounds against a therapeutic target of interest. The combinatorial explosion of binding poses [6,7] and ligand conformations [6,8,9] and the chaotic nature of such dynamical systems [10] prevent popular virtual screening methods from producing safe and effective therapeutic leads a priori. These issues are exacerbated by the fact that virtual screening studies usually consider a single protein target, whereas drugs ingested by humans go through absorption, dispersion, metabolism, and excretion (ADME) and exert their effects

(and side effects or toxicity (T)) via interactions with multiple targets and systems [3,11–15]. Furthermore, the chemical space explored by virtual screening is limited to a relatively small selection of compounds when compared to the vastness of the small molecule space [4,16], thus missing more effective and safer leads.

Computational methods are efficient, accurate, holistic (i.e., take into account the entire interaction space of chemical entities), and have breadth in terms of chemical space exploration necessary to overcome the limitations of traditional approaches [2,6,12,13,17–34]. To expand compound libraries utilized in screening, combinatorial chemistry and machine-learning design pipelines have been developed to generate libraries of compounds likely to bind to a given target [35–37]. Some notable examples in machine learning include Insilico Medicine's Chemistry42 platform, which designs compounds to a binding pocket [38], or a recent transformer-based network that utilized machine translation methods to generate binding ligands for the amino acid sequence of a target protein [39]. However, to take full advantage of these leads, additional screening and in vivo work must be performed to identify off-target binding as these approaches do not address the multitarget nature of drug interactions [11,13].

Various encoder–decoder models [40–42] for conditional [43] molecular generation on multiple properties have been proposed [17,44–46], but in most cases, these properties are limited to physiochemical ones. These models, however, do show great promise in their ability to rapidly generate compounds with desired properties. The most sophisticated conditional molecular generation performed thus far to our knowledge is inducing a differential expression profile of several hundred genes [17]. In all these models, proteins, the functional molecules that are primarily bound by human-ingested drugs to ensure efficacy and ADMET, remain to be considered explicitly on a large scale. Determining interactions between drug candidates and target proteins on a proteomic scale will offer the most comprehensive predictions for bioactivity and safety as many on- and off-targets will be considered simultaneously.

We developed the Computational Analysis of Novel Drug Repurposing Opportunities (CANDO) platform for shotgun multitarget drug discovery, repurposing, and design to overcome the aforementioned limitations of traditional single-target approaches [18–29]. The platform screens and ranks drugs/compounds for every disease/indication (and adverse event) through the large-scale modeling and analytics of the interactions between comprehensive libraries of drugs/compounds and protein structures. CANDO is agonistic to the interaction scoring method used; two primary pipelines within the platform allow for rapid screening and assessment of billions of drug/compound to protein interactions with fast bioanalytic docking and machine learning affinity regression protocols. Machine learning is also used to improve performance in conjunction with preclinical data in an iterative manner. Finally, CANDO implements a variety of benchmarking protocols for shotgun repurposing, i.e., to determine how every known drug is related to every other in the context of the indications/diseases for which they are approved, which enables the evaluation of various pipelines and protocols within and external to the platform for their utility in drug discovery. The multiple fast and accurate interaction scoring/docking protocols, the proteomic scale, and rigorous all-against-all benchmarking used within the platform make it unique and ideal for the design of chemical entities that target a desired proteomic space or objective.

Here, we describe the development and rigorous benchmarking of a multi-step deep learning pipeline for drug design. These pipelines perform conditional drug design given a desired proteomic interaction signature using a generative approach to explore the vastness of the entire small molecule space, while evaluating the functional behavior of candidate designs across the proteomic space. The CANDO platform's benchmarking strategy is used in a modified fashion to determine the performance of the designed compounds relative to an objective drug. We show that the best generated designs were evaluated as being equivalent or better than a variety of controls for twenty objective drugs. Our pipeline represents a significant leap in automating holistic drug design with machine learning,

with the ability to rapidly generate effective and safe drug candidates that accurately target multiple proteins within a proteome as desired.

2. Results and Discussion

2.1. Behavioral Similarity of Designed Compounds to Their Objectives

We observed excellent performance of our Reduced Conditional Variational Autoencoder (RCVAE) proteome-scale design pipeline in all our benchmarking experiments following training (see the methods Section 3.2). If this performance continues to hold following synthesis and validation, it indicates that this pipeline will greatly enhance the pharmaceutical discovery pipeline for novel treatments against a variety of simple and complex indications. A critical aspect of verifying the utility of the designs generated was to compare their predicted behavior (i.e., proteomic interaction signatures) to their intended behavior, which was input to conditional generation. If the predicted behaviors of designed compounds were highly similar to the conditional objective across objectives relative to the corresponding controls, we concluded that the RCVAE design pipeline may be used to accurately design compounds that possess any desirable bioactivity and subsequent function, given the extensive benchmarking and validation the CANDO paradigm has undergone [18,21,26,29,47–50]. This is the primary motivation and goal for using the CVAE architecture in terms of accelerating drug discovery: design with respect to arbitrary numbers of on-, off-, and anti-targets (Figure 1).

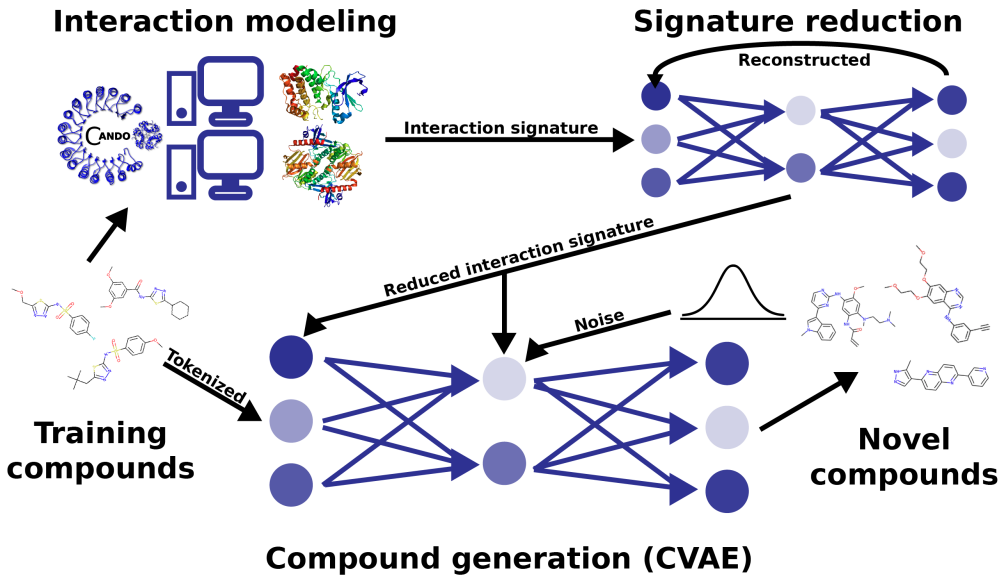

Figure 1. Deep -learning architecture and pipeline for generative drug design. We used the CANDO platform to predict interaction signatures for each compound in a training set against a library of nonredundant protein structures representing the human proteome. The interaction signatures have their dimensionality reduced in an autoencoder, which models the underlying correspondence between protein structures as they behave in the proteome. The reduced signatures are then used as labels for each training compound, which the generative conditional variational autoencoder model learns to reconstruct given a target interaction signature. This pipeline allows us to redesign behaviorally similar compounds to existing drugs based on their interaction signatures, as well as to modulate interactions on a proteomic scale as desired to generate behaviorally novel therapeutics.

We evaluated the performance primarily by the median of each distribution, indicated by the horizontal bars in the box plots in Figure 2. Lower root-mean-squared deviation (RMSD) values indicate a greater reproduction of proteomic interaction signatures for the intended and/or predicted behavior of any given compound, i.e., greater behavioral similarity. Every set of redesigns performed significantly better (p-value of ≤ 0.05) than a selection of random compounds according to this criterion. Additionally, despite being comparatively close in predicted proteomic behavior, our redesigns maintained high levels of structural diversity, as evidenced by their Tanimoto coefficients to our drug library (average ≤ 0.39). For sixteen of the objectives (excluding sirolimus, cucurbitacin Q1, digoxin, and myriocin), the redesigns significantly outperformed the corresponding top 100 and same indication controls. The top 100 control, which included several "me too" compounds (or structural analogs) for each objective [24,51], is the most rigorous one we could devise and illustrates that just generating 100 designs in many instances produces more behaviorally similar compounds to a desired interaction signature than selecting the most similar 100 compounds from a total of 13,194 (the size of the CANDO drug library). The existence of structural analogs in the top100 control indicates bias in favor of already effective compounds in an effort to break into a new market or retain market dominance by generating new intellectual property. New drugs are often derivatives of existing ones with small changes, which our design pipeline is able to overcome, particularly with a bit of extra effort (see Section 2.4 below). Overall, these results indicate that the RCVAE design pipeline produces compounds that accurately match the behavior of desired proteomic interactions relevant to drug discovery. In other words, interactions related to therapeutic efficacy, ADME, and toxicity are modulated precisely in the designed compounds, particularly for objectives from the rational sources subset.

2.2. Relative Performance Gains of Designs Relative to Controls

The bottom panel of Figure 2 compares the relative performance gains for proteomic objectives from the rational design and natural sources subsets relative to the controls. Compounds in the natural sources subset were derived from massively parallel evolutionary processes over eons of time. As a result, they exhibit evolutionary drift, resulting in complex behaviors that are suboptimal from a therapeutic discovery perspective, i.e., unnecessary off-target interactions and/or individual interactions drifting away from functional free energy minima [52]. In addition to verifying that our designs behave as intended, our benchmarking also shows that the RCVAE design pipeline accomplishes this replication of behavioral similarity through an abstraction of rational drug design. That is, the similarities of redesigns to objectives from the rational design subset were greater relative to those from the natural sources subset (Figure 2). Adopting an abstraction of rational drug design is optimal for the impact of this platform on drug discovery because if the model was merely replicating the molecular structure of design objectives and not proteomic behavior specifically, the limitations of natural products (low synthetic accessibility, poor ADMET [53]) would present themselves in the designs in addition to indicating that the model may be over-trained. This discrepancy in the similarity of redesigned natural products and rationally designed drugs, therefore, further supports the notion that the RCVAE pipeline is able to intelligently design compounds with desirable bioactivities across multiple targets.

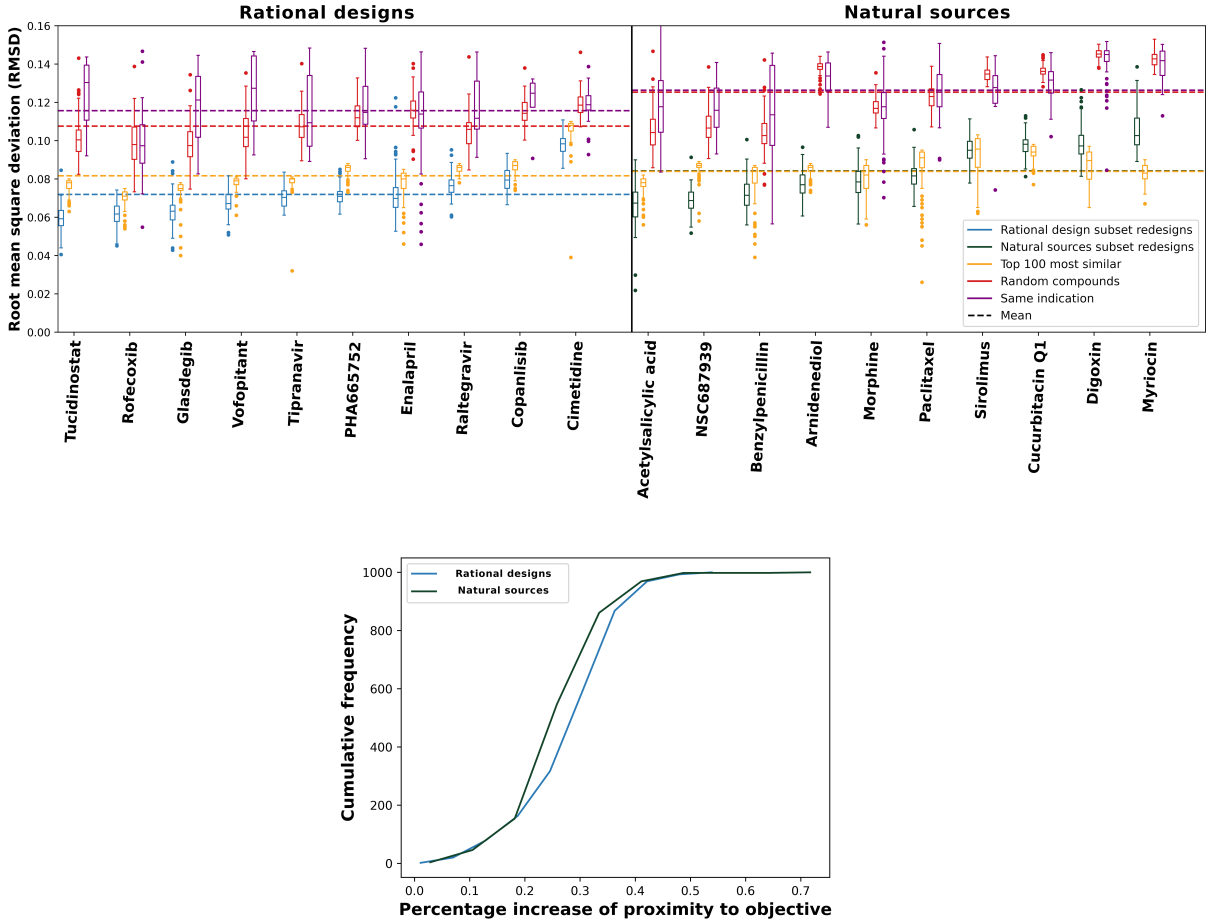

Figure 2. Performance of our deep learning drug design pipeline. To evaluate the performance of the RCVAE pipeline for drug design, we compared redesigned compounds to three controls. The root-mean-squared deviations (RMSDs) between redesigned and control interaction signatures for the rational design and natural sources subsets and the corresponding objective signatures were used to evaluate the performance, as a proxy for behavioral similarity. In the top panels, the blue and green box plots denote the distributions of RMSDs between predicted proteomic interaction signatures for 100 redesigns and the ten corresponding objectives from the rational design (left) and the natural sources (right) subsets along the horizontal axis, respectively. For each RMSD distribution box plot, the boxes indicate the first and third quartile ranges, the horizontal bar indicates the median, and the whiskers indicate non-outlier ranges, with outliers plotted as dots. As a naive control, the red box plots in both panels denote the distributions of the RMSDs between predicted proteomic interaction signatures for a set of 100 randomly selected drug-like compounds taken from the ZINC database [54] and the corresponding objectives. Yellow box plots denote the distributions of the CANDO-predicted top 100 most similar compounds by interaction signature and their corresponding RMSDs; this is a more rigorous control, which checks to see if there exists any compound in the CANDO library that could match or exceed the performance of the design pipeline. Purple box plots denote the RMSD distributions of compounds approved for the same indication as the objective; this is a third control based on phenotype. Dots represent RMSD values for outlier points for each distribution. Dashed lines show the average RMSDs of all compounds corresponding to a given color. All redesign–control distribution pairs were significantly different as indicated by Kolmogorov–Smirnov (K-S) tests and outperformed random controls. For 16/20 of the objective compounds, our redesigns were able to perform better than the top 100 and same indication controls (medians and distributions). The exceptions were four compounds from the natural sources subsets (sirolimus, cucurbitacin Q1, digoxin, and myriocin), discussed further in Sections 2.2 and 2.4. The bottom panel displays cumulative frequency graphs of percentage increases of similarities ("proximity") to the objective for rational design and natural sources subsets redesigns (see the Materials and Methods Section 3.3). Redesigns from the former subset were significantly more accurate to the objective compound than the natural sources subset when compared to the naive control, with mean percentage similarity increases of 33.4% and 32.9%, respectively (p-value $\leq 7.0 \times 10^{-6}$). This indicates that not only does the design pipeline generate compounds with the desired proteomic-scale behavior, but that it likely implements a learned abstraction of rational drug design.

2.3. Visualizing and Filtering Using t-SNE Plots

For all objectives, the redesigns greatly outperformed existing drugs approved for the desired indication when compared to the objective signature. Despite this, some Kolmogorov–Smirnov (K-S) tests indicated weaker statistical significance when differentiating between the RCVAE design and the same indication distributions. To understand the distributions of redesigned and control compounds and to provide an additional filtering mechanism to evaluate the performance of our designs as illustrated in Figure 2, we visualized the interaction signatures of all compounds (objectives, redesigns, and all controls) evaluated in our benchmarking with t-SNE plots [55] (Figure 3).

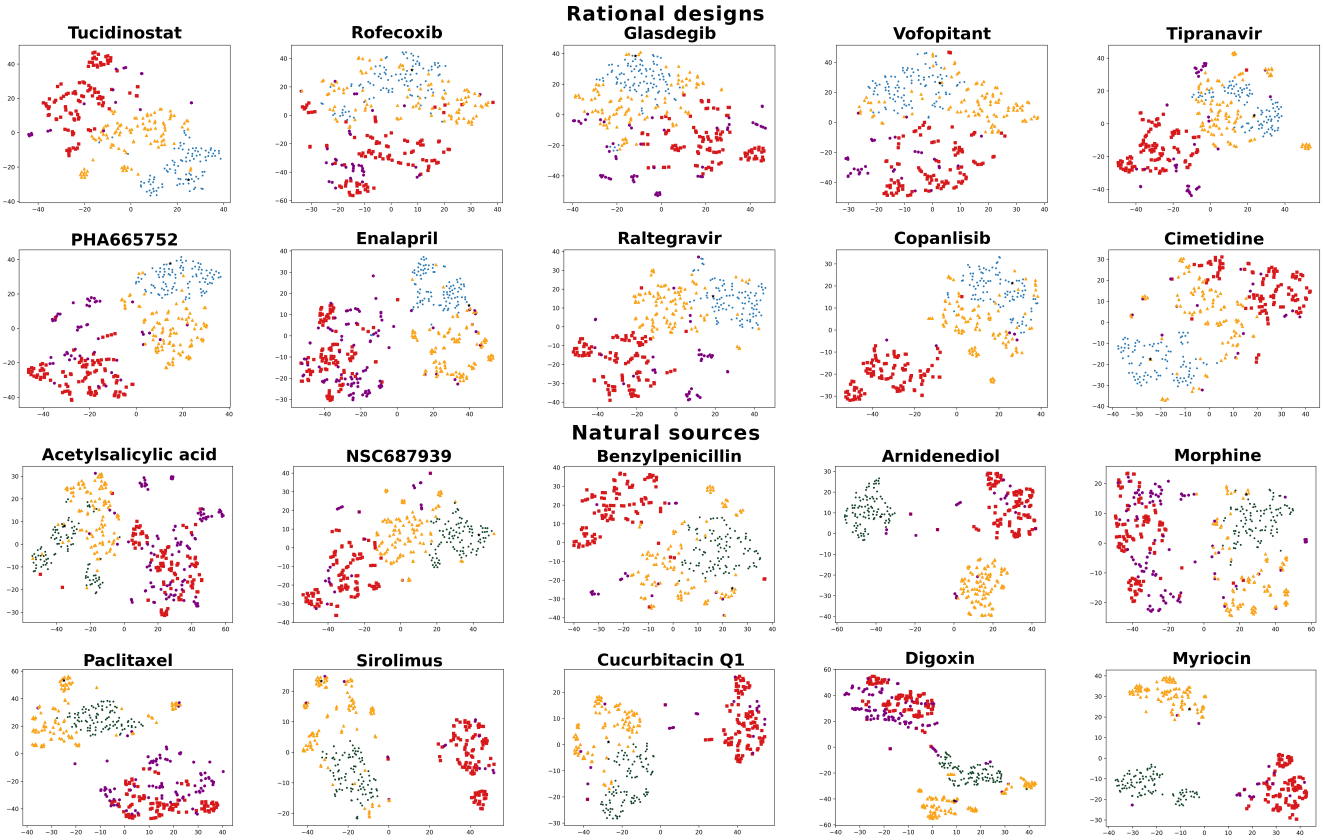

Figure 3. t-SNE visualizations for the interaction signatures of objective compounds, redesigns, and various controls with structural comparisons. t-SNE plots were generated for each of the twenty design objectives. The t-SNE algorithm was run on the interaction signatures of the objectives (black stars) and 100 redesigns (blue and green circles), as well as the 100 random (red squares), top 100 (orange triangles), and same indication (purple hexagons) control compounds. (Color coding is the same as in Figure 2.) Euclidean distances shown in t-SNE visualizations generally corroborate our findings from Figure 2 when considering the RMSD due to the maintenance of proximal points and distributions. The exceptions to this were for benzylpenicillin and paclitaxel, as Figure 2 shows greater proximity for designs than the top 100 control, whereas t-SNE plots show better clustering for the top 100 set around the objective. The generally (14/20) greater proximity of designed compound clusters to the objective point when compared to top 100, same indication, and random control compounds corroborate the behavioral similarity of designed compounds to their objective in the predicted interaction space, i.e., designed compounds are predicted to behave in the manner in which they were designed.

t-SNE plots generally corroborate the relative behavioral similarities between redesigns and controls relative to their objectives. RCVAE pipeline designs tend to cluster densely around the objective compound, as they are designed to do. The top 100 compounds cluster around these designs, with a few structural analogs very close to their objectives. Finally, the same indication and random compounds clustered the farthest from the objective. As the same indication compounds represent a group of diverse drugs approved for an indication that includes the objective, there is a general lack of clustering,

and they are at greater distances from their objectives. Comparing the performance of the designs relative to their objectives and controls using both Figures 2 and 3 indicates that the t-SNE plots illustrated in the latter may be used to assess the confidence in and room for improvement of the design pipeline performance for specific objectives.

2.4. Improving Cases with Sub-Optimal Performance

For eleven objectives, a handful of outliers in the top 100 or same indication controls outperformed the design distributions (Figures 2 and 3). As noted above, structural analogs create a bias when evaluating performance due to them having very similar behavioral interaction signatures. We observed that the top 10 compounds (out of the top 100 controls, covering almost all outliers) yielded an average Tanimoto coefficient of 0.51, in contrast to an average of 0.39 for all designs, relative to their objectives. We further observed that 23/200 top 10 compounds, across all 20 objectives, had a Tanimoto coefficient $\geq$0.90 with an average of 0.96, demonstrating the "me too" bias with the top 100 control. Unlike the few top 100 outliers, the designs offer a larger and more diverse selection pool for prospective validation.

Regardless, we further investigated the behavior of the RCVAE pipeline for one of the objectives (cucurbitacin Q1) where an outlier was clearly better than the best designed compound by expanding the number of designs generated by the RCVAE pipeline from 100 to 1000 compounds. We found that the RMSDs of the top 100 out of 1000 designs, far less than the 13,194 compounds that were the source for the top 100 most similar compounds control, ranged from 0.073 to 0.09. This placed the RMSD of the best designs well below the lowest RMSD outlier of the top 100 control. Altogether, this indicates that the performance of the RCVAE pipeline may be enhanced by increasing the number of designs generated and selecting for behavioral similarity. Our design pipeline offers structurally diverse lead compounds with the potential to match or exceed the behavioral similarity of the top CANDO predictions from its known drug library for noisy objectives, such as compounds from the natural sources subset.

2.5. Synthetic Feasibility of Designed Compounds

To viably demonstrate that our 2000 redesigns were synthetically feasible, we utilized a high-throughput machine-learning-based approach to predict synthetic complexity scores, called SCScore. SCScore utilizes a database of known synthetic reaction pathways for training to make predictions of how easy or difficult it is to synthesize a novel compound.

As shown in Figure 4, predicted synthetic complexities for the RCVAE pipeline redesigns are often comparable to their design objective, i.e., the objective compound's scores exist within the distribution of the scores for the redesigns. Low Tanimoto coefficients (average 0.39) between designs and objective compounds indicate that the comparable synthetic complexities of our redesigns are not due to a corresponding high structural similarity. In other words, despite the high structural diversity of the generated designs, the synthetic complexity for objective and designed compounds remains stable. This may be explained by the maintenance of functionally relevant substructures of comparable synthetic complexity that are present in any given objective and redesign that are combined differently to produce similar behaviors and low structural similarity. We are studying this phenomenon more thoroughly with larger datasets by investigating the redundancy of the substructures of all approved drugs and redesigns for publication in a future study.

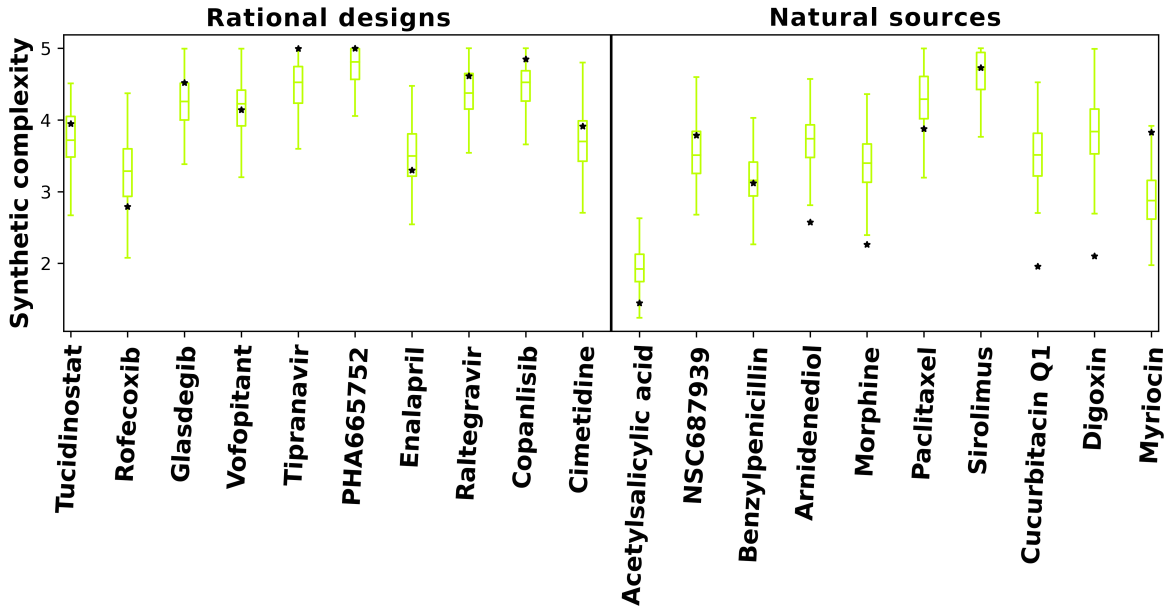

Figure 4. Predicted synthetic complexity of designs compared to objective compounds. Synthetic complexity scores for objective and RCVAE pipeline designed compounds as predicted by SCScore [56] serve as the basis for this comparison. Box plots denote the distributions of predicted synthetic complexity scores for RCVAE redesigns of objective compounds (stars). The scores of the objectives enable us to evaluate the synthetic feasibility of the redesigns relative to their corresponding objective compounds. There is a greater occurrence of low-scoring natural sources objectives in predicted synthetic complexity. This indicates that the SCScore software is biased by what is already well known or it may also highlight a potential reason why RCVAE performs better for objectives from the rational design subset: it is somewhat plausible that evolution optimizes for synthetic accessibility over binding free energies, hence the discrepancies in behavioral similarities between the designs and their objectives for the two benchmarking (rational design and natural sources) subsets (Figure 2). RCVAE designs are of similar synthetic complexity to their corresponding objective compounds as most objective complexities lie within the box plot whiskers (not outliers), with some designs being more accessible than their objectives.

Additionally, predicted synthetic complexities are typically lower for objectives and redesigns in the natural sources subset relative to the rational design one, indicating that the SCScore software may be biased by what is already well known. As objectives in the natural sources subset score similarly to their redesigns, it is somewhat plausible that evolutionary optimization towards synthetic accessibility, at the expense of macromolecular interaction free energies [52], may account for the diminished behavioral similarity of RCVAE designs between objectives in the rational design and natural sources subsets correspondingly (see Figure 2). A comprehensive analysis of this hypothesis with larger subsets is necessary to validate or falsify this hypothesis.

It is useful to note that many RCVAE designs are more synthetically accessible than the compound whose behaviors they replicate. As Figure 2 already indicates high behavioral similarity between redesigns and their objectives, the RCVAE design pipeline may serve the additional purpose of designing analogs to existing drugs that are more synthetically feasible and therefore easier and less costly to produce. This is accomplished without adding a synthetic complexity parameter to the condition vector for compound generation (Figure 4).

Finally, we routinely used several other methods to evaluate RCVAE designs such as the Quantitative Estimation of Drug-likeness (QED) [57], Synthetic Accessibility Score (SAScore) [58], and AiZynthFinder [59] to assess their chemical viability and drug-likeness. We also compared our designs to benchmarks from GuacaMol [60]. These results corroborated the outputs of our benchmarking and/or SCScore; future work will include a rigorous evaluation of drug design technologies, much as we have done for repurposing [29].

2.6. Applications to Aging and Non-Small Cell Lung Cancer

A fundamental tenet of CANDO is that evaluating all the possible interactions between a human-ingested drug/compound and the macromolecules and systems it encounters on a proteomic/interactomic scale is necessary to determine its safety and efficacy for a given indication [18–29]. The RCVAE design pipeline represents a significant step forward in early drug discovery as it allows for the virtually unlimited generation of novel putative drug candidates to treat any indication/disease by combating it on a proteomic scale. As a precursor to upcoming work, we generated designs for several objective compounds/drugs associated with human or cell longevity ("aging") [61–70] and approved for Non-Small Cell Lung Cancer (NSCLC) [71–78]. We expect our pipeline to produce redesigns that retain the proteome-scale behaviors of these objectives used to treat these complex indications/diseases, while being structurally diverse.

Figure 5 illustrates the top redesigns for the thirteen objectives covering the two indications ranked using two metrics based on the greatest similarity criterion: lowest RMSD between corresponding interaction signatures and highest Tanimoto coefficient between corresponding molecular fingerprints. Two classes of redesigns, all with the greatest behavioral similarity to their objectives, emerged when performing the comparisons illustrated in Figure 5: designs that chemically/structurally resemble their objective compound/indication (metformin, NAD+, resveratrol, curcumin, and RepSox for aging and gefitinib, erlotinib, afatinib, and dacomitinib for NSCLC, all with a Tanimoto coefficient $\geq$0.39) and those that do not. The former class is intriguing as it implies the existence of highly optimized structures for a given phenotype (indication), as shown by the convergence of redesigns to known drugs. It also demonstrates that the RCVAE design pipeline produces highly similar designs to known drugs to perform specific tasks, only from their proteomic interaction behavior and without being exposed to any similar structures in training. In other words, the proteomic-scale interaction information for a compound is enough information for our design pipeline to reliably reconstruct a chemically/structurally similar compound in some cases. The latter class demonstrates the expanse and diversity of a chemical space not yet charted by medicinal chemists, capable of replicating the interactions of known drugs. We are in the process of synthesizing the top designs from these pipelines for these indications and validating them in corresponding preclinical models via industry partners and collaborators.

2.7. Limitations and Future Work

To treat an indication in a comprehensive fashion, especially complex ones such as aging or NSCLC, entirely novel drugs will likely need to be developed that go beyond replicating the systemic effects of existing ones. This would require a thorough and accurate description of the interaction networks responsible for disease etiology, as well as compound behavior to ensure optimal efficacy and safety. The creation of these interaction networks may be accomplished by multiscale modeling, literature/database analyses, and/or high-throughput experimental studies, all of which may be incorporated within CANDO. We are currently in the process of conditioning the RCVAE design pipelines and comparing them to ones based on graph neural networks [79] using these more complex interaction networks that go beyond the information present in the linear signatures.

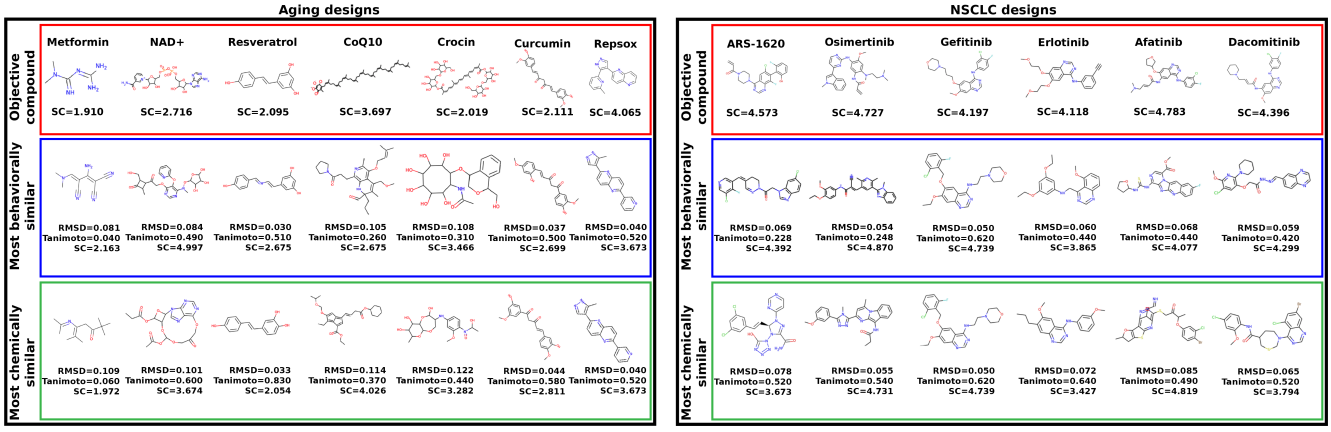

Figure 5. Analysis of the top designs generated for aging and Non-Small Cell Lung Cancer (NSCLC). The names, chemical structures, and predicted Synthetic Complexity (SC) scores of the objective compounds for aging (left) and NSCLC (right) are displayed in the top row of each panel (red border). The chemical structures of the most similar redesigns to each objective based on their interaction signatures alongside three metrics for these redesigned compounds relative to their objectives (RMSD of interaction signatures, Tanimoto coefficient, and predicted synthetic complexity) are displayed in the row below (blue border). The chemical structures of the most similar redesigns using molecular fingerprints alongside the same three metrics (RMSD, Tanimoto, and synthetic complexity) for these redesigns relative to their objectives are displayed in the bottom row (green border). The designs generated using the RCVAE pipeline retain a fair amount of structural diversity, as indicated by the fingerprint comparison scores (low Tanimoto coefficients) whilst maintaining high predicted behavioral similarity in terms of interaction signatures (low RMSDs). Other methods for evaluating our designs such as QED corroborated the above results [57]. These designs demonstrate the utility of our pipeline for designing novel compounds to combat complex indications on a proteomic scale and may be pursued further via preclinical validation studies.

We are currently exploring other scoring protocols within the CANDO platform for conditional generation to overcome the limitations of any specific interaction calculation method. For example, the interaction scoring protocol used in this work has been shown to have great utility in the context of evaluating proteomic behavioral similarity based on benchmarking performance [18–29]. However, information on agonism/antagonism, and downstream functional activity upon binding may be obtained via design pipelines that utilize gene and protein expression data, which we are incorporating into CANDO. Public gene expression data, available through the L1000 and Connectivity Map projects [80,81], highlight important genes/proteins that are upregulated and downregulated following exposure to a drug/compound. For example, if gene expression data suggest that certain downstream genes are significantly upregulated in a given pathway following interaction with a compound and that same compound is predicted to have strong interactions with multiple proteins in that pathway, it can be inferred that those are likely to cause activating/agonist behavior. The same can be inferred for downregulated genes and inhibition/antagonism. In addition, we are exploring the use of high-throughput robotic systems such as DESI-MS to generate large-scale interaction and activity data [82–90]. CANDO thus enables and illustrates the benefit of combining heterogeneous sources (gene, as well as protein expression, protein pathway databases, high-throughput binding, and activity data) to create novel types of interaction signatures/networks to produce design objectives that tackle complex indications.

The CANDO platform enables the benchmarking of any arbitrary proteome/protein library for its utility in drug discovery using a similar all-against-all process as described in Section 3.3. The generation of highly accurate modeled protein structures such as those predicted by Deepmind's AlphaFold offer an attractive representation of the full human proteome to perform such benchmarking, which we have completed and will publish separately. This allows for conditional drug design using AlphaFold interaction signatures, especially giving us greater coverage and control over particular proteins and pathways to modulate with expert input for specific indications.

The benchmarking and performance evaluation of our RCVAE drug design pipeline were based on using known data as the ground truth or gold standard. While computational experiments are an important first step and indicate promising, prospective preclinical validation of the pipeline, its designs will require medicinal chemistry synthesis, binding studies, and disease models assays at multiple scales, which we are currently undertaking. Regardless, our combined work to date [18–29], including this study, indicates that novel high-throughput methods for rapidly identifying relationships between compounds, proteins, pathways, and cells are highly desirable for holistic drug discovery. As CANDO is agnostic to the specific methods used for any of its protocols, should such data become available, the RCVAE design pipeline described here would be well poised to take advantage of them for maximum drug discovery efficiency.

3. Materials and Methods

Figure 1 illustrates our overall methodology to create a new drug design pipeline. We employed the Computational Analysis of Novel Drug Opportunities (CANDO) platform to generate proteomic interaction signatures for the compounds in the training set of our learning-based model. The CANDO interaction signatures had their dimensionality reduced in an autoencoder, which models the underlying correspondence between protein structures as they function in the proteome. The reduced signatures were then used as labels for each training molecule, which the generative Conditional Variational Autoencoder (CVAE) model learns to reconstruct given a target interaction signature. We then used CANDO to benchmark the performance of the designed compounds in the context of their objectives and make predictions of novel designs for two indications for future prospective validation.

3.1. Compound–Proteome Interaction Signature Generation Using the CANDO Platform

Multiple pipelines for multiscale therapeutic discovery, repurposing, and design have been implemented in the CANDO platform [18–29]. Here, we utilized CANDO to simulate the interactions between a given drug/compound and a library of protein structures to generate the corresponding proteomic interaction signature.

The protein structure library used in this study is a set of 14,606 nonredundant structures derived from the Protein Data Bank (PDB) [91] corresponding to the human proteome fold space ("nrPDB")[92–97]. The compound–protein interaction scores in these signatures are computed using the bioanalytic docking protocol BANDOCK, which compares query compound structures to all ligands that are known or predicted to interact with a protein binding site [22,27,29].

Potential binding sites on a protein are elucidated using the COACH algorithm, which uses three different complementary algorithms and a consensus approach to consider the sequence or substructure similarity to known PDB binding sites [98]. COACH has been utilized extensively within the CANDO platform to accurately predict the binding behavior of numerous compounds against numerous targets, as demonstrated by its benchmarking performance in multiple studies (Section 3.3 and [18–29]). For each potential binding site, the COACH output includes a set of co-crystallized ligands, which are compared to a compound of interest using binary chemical fingerprinting methods that describe the presence or absence of particular molecular substructures [99]. The maximum Tanimoto coefficient between the binary fingerprints of the query compound and the set of all predicted protein binding site ligands becomes the interaction score. The better the score, the higher the likelihood of the interaction being correct due to the inferred homology. Thus, if there are proteins with multiple binding sites and corresponding ligands, the strongest interaction is used. If there are no matches, then the score returned is zero (i.e., no interaction). The final output is a vector of 14,606 scores comprising the interaction signature between a given compound and the nrPDB library. Further detail on the pipelines used to generate and benchmark the interaction signatures is given elsewhere in numerous publications [18–29].

3.2. Model Architecture and Data Generation

We selected a CVAE [45] architecture for generating novel molecular structures. Training data consisted of SMILES [100] strings labeled with predicted proteomic interaction signatures based on the nrPDB library. The training set consisted of 300,000 compounds selected at random from the ZINC database [54]. The 14,606 protein binding scores were predicted for each compound. The dimensionality of the protein interaction signatures was reduced to a 200-dimensional vector via a conventional autoencoder [101]. Following an input layer with 14,606 neurons, the encoder consisted of 10 sequential, densely connected layers with 10,000, 7750, 5500, 2250, 2000, 1250, 1000, 500, 250, and 200 neurons in each layer, respectively. This was reversed in the decoder, and a final layer with 14,606 neurons was used as the output to the network. The root-mean-squared deviation (RMSD) between the input and reconstructed signatures was used as a loss metric. This model was trained on 250,000 compounds until over-fitting was observed, which occurred after 15 epochs/iterations. Each epoch was validated on another 50,000 randomly selected compounds. The model was then used to reduce the non-redundant signatures of the training compounds. These became the labels for each SMILES string present in the CVAE training data.

Before being input into the CVAE, SMILES strings were one-hot encoded [102,103], resulting in a rank-2 tensor of size (sequence length $\times$ vocab length), where sequence length is the maximum number of characters allowed per SMILES string and vocab length is the unique number of characters represented in the input data. The reduced protein signature, c, was appended to the end of the one-hot encoding repeated at each sequence position (commonly referred to as time steps in the context of Long Short-Term Memory (LSTM) cells [102,104,105]). Similar to [45], the tensor was fed sequentially through three LSTM cells [104,105] to encode the original input. The encoder outputs to two parallel layers, one representing the mean and one for the standard deviation of the latent vector. The latent vector, z, consists of 200 dimensions and is sampled from the encoder output. The latent vector is then repeated for the total number of time steps, and the protein signature, c, is re-appended onto the resulting tensor in the prior fashion. This is input into the decoder, which also consists of three LSTM cells. Finally, this is output to a matrix of probabilities for each character in a SMILES string. Taking the maximum probability token for each character slot, one-hot encoded SMILES strings denoting reconstructions of input compounds are generated.

The loss metric used to train the CVAE is as follows:

$$E[log(P(X|z,c))] - D_{KL}[Q(z|X,c)||P(X|z,c)] \quad (1)$$

where E denotes the reconstruction error and D_{KL} denotes the relative entropy or the Kullback–Leibler divergence [106]. $P(X|z,c)$ denotes the probability density function approximated by the decoder for each character in a SMILES string given the latent and conditional vectors. $Q(z|X,c)$ denotes the probability density function approximated by the encoder given the input SMILES strings and condition vector. The CVAE was trained on 300,000 compounds until convergence. The Reduced CVAE (RCVAE) model pipeline is depicted visually in Figure 1.

3.3. Benchmarking and Analysis of the RCVAE Design Pipeline Performance

The fundamental supposition and result of benchmarking the CANDO platform is that compounds with similar interaction signatures will behave similarly. On a proteomic scale, these behaviors take the form of efficacy and ADMET for a given indication. The CANDO platform was benchmarked using known drug-indication associations [18–29] derived from the Comparative Toxicogenomics Database [107] and, more recently, drug-adverse events obtained from OFFSIDES [108,109] and SIDER [109]. We recently published the best metrics to use for benchmarking drug repurposing platforms [29]. The results of benchmarking and prospectively validating CANDO indicate that the proteomic-scale interaction modeling of drugs elucidates their behaviors, and these behaviors correspond

to treatments for indications for which these drugs are approved [18,21,26,29,47–50]. The benchmarking of the platform using known associations in this comprehensive all-against-all manner enables us to assess the correctness and utility of other parameters, such as the protein library composition, solved vs. modeled structures, different molecular docking and machine-learning algorithms, etc.

We performed several benchmarks of the RCVAE to verify its utility and robustness beyond that of its performance based on metrics used in training. Our benchmark set consisted of twenty approved or experimental objective drugs, comprised of subsets of ten derived from rational design and natural sources, respectively (Figure 2). These compounds and all related ones with a Tanimoto [110] coefficient $\geq$0.9 were omitted from training. Proteomic interaction signatures were computed by CANDO, and output SMILES strings were generated by the RCVAE design pipeline as described above. This resulted in SMILES strings corresponding to 100 novel redesigns for each objective drug. One-hundred compounds were selected at random from the curated ZINC database to serve as a naive control. As a second, more rigorous control, the CANDO platform was used to generate the 100 most similar compounds ("top 100") from its 13,194-sized library to each objective drug according to their proteomic interaction signatures. As a third control, the CANDO platform's indication prediction pipeline was used to predict a set of compounds for the indication associated with each objective, i.e., the indication that the objective is approved for ("same indication") [27,107]. Proteomic interaction signatures were generated for all compounds (designs and controls), which were then compared to that of the objective using the RMSDs between them. The RMSD distributions are illustrated using box plots with boxes depicting the first and third quartile ranges, horizontal bars depicting the median, whiskers representing non-outlier ranges, and outliers explicitly plotted (Figure 2, Results Section 2.1).

Kolmogorov–Smirnov (K-S) tests [111,112] were used to demonstrate statistical significance [113] between samples for each redesign–control RMSD distribution pair for each objective. We also compared the performance of the RCVAE design pipeline between objectives from the rational design and natural sources subsets, respectively. To do this, we computed the average RMSD between each naive control and the corresponding objective and compared this to the RMSD of each redesign for an objective. This yielded a percent increase of similarities to the objective for each redesign given by:

$$P = \frac{< RMSD_{control} > - RMSD_{redesign}}{< RMSD_{control} >} \quad (2)$$

where P denotes the percent increase ("proximity"), $< RMSD_{control} >$ denotes the average RMSD between the naive control and corresponding objective, and $RMSD_{redesign}$ denotes the RMSD of the redesign when compared to the objective proteomic interaction signature. This yielded 1000 total values of percent proximity increases for both the rational design and natural sources subsets. The values for both subsets were then averaged and compared using K-S tests (Figure 2, Results Section 2.1).

To better visualize the distributions of our redesigns and controls, t-distributed Stochastic Network Embedding (t-SNE) plots [55] that show the clustering of similar interaction signatures in two dimensions were generated with the interaction signatures of all objective, redesign, top 100, and same indication compounds (Figure 3, results Section 2.3).

We also computed Tanimoto coefficients for all redesigns relative to their respective objective compounds to determine structural diversity. Finally, we utilized SCScore [56], a machine-learning platform to predict the synthetic complexity of our redesigns in relation to the corresponding objectives (Figure 4, Results Section 2.5).

3.4. Generating Novel Designs for Prospective Validation

Design objectives for benchmarking were selected from a diverse set of indications and approved/experimental statuses to ensure broad coverage of the proteomic interaction signature space and to mitigate potential bias in the results. Regardless, the benchmark

set was repeatedly used to parameterize the pipeline described here, which has the potential to lead to overtraining. To address this and also to apply our design pipelines to relevant real-world problems of sufficient complexity where the proteomic approach would be relevant, we selected 13 (7 + 6) objective compounds that have shown promise for aging/developmental intervention [114–118] and NSCLC [118–120] to redesign for prospective validation (Figure 5, Results Section 2.6).

4. Conclusions

We utilized the RCVAE pipeline within the CANDO platform to take advantage of multiscale compound–proteome interaction modeling and develop an attractive approach to holistic drug design. We compared the predicted behaviors of the designed compounds to those of known drugs/compounds and demonstrated that the RCVAE pipeline is capable of generating novel compounds with the desired specificity of binding on a proteomic scale. We additionally demonstrated that compounds designed by our pipeline maintained reasonable predicted synthetic complexities and were structurally diverse. We expect the compounds designed using our pipeline for aging/developmental intervention and NSCLC will serve as novel leads for safe and effective therapeutics following prospective validation. The RCVAE design pipeline generates novel compounds that are synthetically feasible and behaviorally desirable, simultaneously taking efficacy and ADMET into account by examining interactions on a proteomic scale, which is necessary to understand the science of small molecule behavior and apply it to holistic therapeutic discovery.

Author Contributions: B.O. conceived of the RCVAE pipeline, research design, approach, and methods, conducted all experiments and analyses, and drafted the manuscript. Z.F. helped with the research design, approach, and methods and editing and proofing of the manuscript. W.M. helped with the research design, approach, and methods and editing and proofing of the manuscript. R.S. conceived of the research design, approach, and methods, supervised the overall project, and edited the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by National Institutes of Health Director's Pioneer Award (DP1OD006779), National Institutes of Health Clinical and Translational Sciences Award (UL1TR001412), NIH T15 Award (T15LM012495), NCATS ASPIRE Design Challenge Award, NCATS ASPIRE Reduction-to- Practice Award, and startup funds from the Department of Biomedical Informatics at the University at Buffalo.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All relevant data are presented/illustrated in this manuscript.

Acknowledgments: The authors would like to acknowledge the support provided by the Center for Computational Research at the University at Buffalo. In addition to all members of the Samudrala Computational Biology Group, we give special thanks to Liana Bruggemann for providing us with the list of NSCLC objectives and Mira Moukheiber for assistance with proofreading.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; nor in the decision to publish the results. The authors have formed multiple startups that seek to commercialize the outputs of the CANDO platform.

References

1. Schuhmacher, A.; Gassmann, O.; Hinder, M. Changing R&D models in research-based pharmaceutical companies. *J. Transl. Med.* **2016**, *14*, 105.
2. Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **2002**, *1*, 882–894. [CrossRef] [PubMed]
3. Dhasmana, A.; Raza, S.; Jahan, R.; Lohani, M.; Arif, J.M. *Chapter 19-High-Throughput Virtual Screening (HTVS) of Natural Compounds and Exploration of Their Biomolecular Mechanisms: An In Silico Approach*; Academic Press: Cambridge, MA, USA 2019; pp. 523–548.
4. Graff, D.E.; Shakhnovich, E.I.; Coley, C.W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci.* **2021**, *12*, 7866–7881. [CrossRef]

5. Lim, J.; Ryu, S.; Park, K.; Choe, Y.J.; Ham, J.; Kim, W.Y. Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation. *J. Chem. Inf. Model.* **2019**, *59*, 3981–3988. [CrossRef] [PubMed]
6. Zoete, V.; Grosdidier, A.; Michielin, O. Docking, virtual high throughput screening and in silico fragment-based drug design. *J. Cell. Mol. Med.* **2009**, *13*, 238–248. [CrossRef]
7. Ha, E.; Lwin, C.; Durrant, J. LigGrep: A tool for filtering docked poses to improve virtual-screening hit rates. *J. Cheminform.* **2020**, *12*, 69. [CrossRef]
8. Lee, H.S.; Choi, J.; Kufareva, I.; Abagyan, R.; Filikov, A.; Yang, Y.; Yoon, S. Optimization of high throughput virtual screening by combining shape-matching and docking methods. *J. Chem. Inf. Model.* **2008**, *48*, 489–497. [CrossRef]
9. Corbeil, C.R.; Moitessier, N. Docking ligands into flexible and solvated macromolecules. 3. Impact of input ligand conformation, protein flexibility, and water molecules on the accuracy of docking programs. *J. Chem. Inf. Model.* **2009**, *49*, 997–1009. [CrossRef]
10. Feher, M.; Williams, C. Numerical errors and chaotic behavior in docking simulations. *J. Chem. Inf. Model.* **2012**, *52*, 724–738. [CrossRef] [PubMed]
11. Boran, A.; Iyengar, R. Systems approaches to polypharmacology and drug discovery. *Curr. Opin. Drug Discov. Dev.* **2010**, *13*, 297–309.
12. Peyvandipour, A.; Saberian, N.; Shafi, A.; Donato, M.; Draghici, S. A novel computational approach for drug repurposing using systems biology. *Bioinformatics* **2018**, *34*, 2817–2825. [CrossRef]
13. Shafi, A.; Nguyen, T.; Peyvandipour, A.; Nguyen, H.; Draghici, S. A multi-cohort and multi-omics meta-analysis framework to identify network-based Gene signatures. *Front. Genet.* **2019**, *10*, 159. [CrossRef] [PubMed]
14. Tatonetti, N.P.; Liu, T.; Altman, R.B. Predicting drug side-effects by chemical systems biology. *Genome Biol.* **2009**, *10*, 238. [CrossRef]
15. Liu, T.; Altman, R.B. Relating essential proteins to drug side-effects using canonical component analysis: A structure-based approach. *J. Chem. Inf. Model.* **2015**, *55*, 1483–1494. [CrossRef]
16. Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432*, 832. [CrossRef]
17. Shayakhmetov, R.; Kuznetsov, M.; Zhebrak, A.; Kadurin, A.; Nikolenko, S.; Aliper, A.; Polykovskiy, D. Molecular generation for desired transcriptome changes with adversarial autoencoders. *Front. Pharmacol.* **2020**, *11*, 269. [CrossRef]
18. Minie, M.; Sethi, G.; Chopra, G.; Horst, J.; Roy, A.; White, G.; Samudrala, R. CANDO and the infinite drug discovery frontier. *Drug Discov. Today* **2014**, *19*, 1353–1363. [CrossRef]
19. Sethi, G.; Chopra, G.; Samudrala, R. Multiscale modelling of relationships between protein classes and drug behavior across all diseases using the CANDO platform. *Mini Rev. Med. Chem.* **2015**, *15*, 705–717. [CrossRef] [PubMed]
20. Chopra, G.; Samudrala, R. Exploring polypharmacology in drug discovery and repurposing using the CANDO platform. *Curr. Pharm. Des.* **2016**, *22*, 3109–3123. [CrossRef]
21. Chopra, G.; Kaushik, S.; Elkin, P.; Samudrala, R. Combating Ebola with repurposed therapeutics using the CANDO platform. *Molecules* **2016**, *21*, 1537. [CrossRef] [PubMed]
22. Mangione, W.; Samudrala, R. Identifying protein features responsible for improved drug repurposing accuracies using the CANDO platform: Implications for drug design. *Molecules* **2019**, *24*, 167. [CrossRef] [PubMed]
23. Falls, Z.; Mangione, W.; Schuler, J.; Samudrala, R. Exploration of interaction scoring criteria in the CANDO platform. *BMC Bioinform.* **2019**, *12*, 318. [CrossRef] [PubMed]
24. Schuler, J.; Samudrala, R. Fingerprinting CANDO: Increased accuracy with structure and ligand based shotgun drug repurposing. *ACS Omega* **2019**, *4*, 17393–17403. [CrossRef] [PubMed]
25. Fine, J.; Lacker, R.; Samudrala, R.; Chopra, G. Computational chemoproteomics to understand the role of selected psychoactives in treating mental health disorders. *Sci. Rep.* **2019**, *9*, 13155. [CrossRef] [PubMed]
26. Mangione, W.; Falls, Z.; Melendy, T.; Chopra, G.; Samudrala, R. Shotgun drug repurposing biotechnology to tackle epidemics and pandemics. *Drug Discov. Today* **2020**, *25*, 1126–1129. [CrossRef]
27. Mangione, W.; Falls, Z.; Chopra, G.; Samudrala, R. cando.py: Open source software for analyzing large scale drug-protein-disease data. *J. Chem. Inf. Model.* **2020**, *60*, 4131–4136. [CrossRef]
28. Hudson, M.; Samudrala, R. Multiscale virtual screening optimization for shotgun drug repurposing using the CANDO platform. *Molecules* **2021**, *26*, 2581–2597. [CrossRef] [PubMed]
29. Schuler, J.; Falls, Z.; Mangione, W.; Hudson, M.; Bruggemann, L.; Samdurala, R. Evaluating performance of drug repurposing technologies. *Drug Discov. Today* **2021**, in press. . [CrossRef]
30. Yang, L.; Wang, K.J.; Wang, L.S.; Jegga, A.G.; Qin, S.Y.; He, G.; Chen, J.; Xiao, Y.; He, L. Chemical-protein interactome and its application in off-target identification. *Interdiscip. Sci. Comput. Life Sci.* **2011**, *3*, 22–30. [CrossRef]
31. Liu, T.; Tang, G.; Capriotti, E. Comparative modeling: The state of the art and protein drug target structure prediction. *Comb. Chem. High Throughput Screen.* **2011**, *14*, 532–547. [CrossRef]
32. Wu, C.; Gudivada, R.C.; Aronow, B.J.; Jegga, A.G. Computational drug repositioning through heterogeneous network clustering. *BMC Syst. Biol.* **2013**, *7*, S6. [CrossRef] [PubMed]
33. Yella, J.; Yaddanapudi, S.; Wang, Y.; Jegga, A. Changing trends in computational drug repositioning. *Pharmaceuticals* **2018**, *11*, 57. [CrossRef]
34. Wang, Y.; Yella, J.; Jegga, A.G. Transcriptomic data mining and repurposing for computational drug discovery. In *Methods in Molecular Biology*; Springer: New York, NY, USA, 2018; pp. 73–95.

35. Patel, L.; Shukla, T.; Huang, X.; Ussery, D.W.; Wang, S. Machine learning methods in drug discovery. *Molecules* **2020**, *25*, 5277. [CrossRef] [PubMed]
36. Yuan, Y.; Pei, J.; Lai, L. LigBuilder V3: A multi-target de novo drug design approach. *Front. Chem.* **2020**, *8*, 142. [CrossRef]
37. Bai, Q.; Tan, S.; Xu, T.; Liu, H.; Huang, J.; Yao, X. MolAICal: A soft tool for 3D drug design of protein targets by artificial intelligence and classical algorithm. *Brief. Bioinform.* **2020**, *22*, bbaa161. [CrossRef]
38. Chemistry 42. Available online: https://insilico.com/chemistry42 (accessed on 30 July 2021).
39. Grechishnikova, D. Transformer neural network for protein-specific de novo drug generation as a machine translation problem. *Sci. Rep.* **2021**, *11*, 321. [CrossRef]
40. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
41. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2014**, arXiv:1312.6114.
42. Polykovskiy, D.; Vetrov, D. Deterministic Decoding for Discrete Data in Variational Autoencoders. In Proceedings of the Machine Learning Research, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Palermo, Italy, 3–5 June 2020; Chiappa, S., Calandra, R., Eds.; 2020; Volume 108, pp. 3046–3056.
43. Sohn, K.; Lee, H.; Yan, X. Learning Structured Output Representation using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
44. Gómez-Bombarelli, R.; Wei, J.N.; Duvenaud, D.; Hernández-Lobato, J.M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T.D.; Adams, R.P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276. [CrossRef] [PubMed]
45. Lim, J.; Ryu, S.; Kim, J.; Kim, W. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *J. Cheminform.* **2018**, *10*, 31. [CrossRef]
46. Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; et al. Molecular Sets (MOSES): A benchmarking platform for molecular generation models. *Front. Pharmacol.* **2020**, *11*, 1931. [CrossRef]
47. Jenwitheesuk, E.; Samudrala, R. Identification of potential multitarget antimalarial drugs. *J. Am. Med. Assoc.* **2005**, *294*, 1490–1491.
48. Jenwitheesuk, E.; Samudrala, R. New paradigms for drug discovery: Computational multitarget screening. *Trends Pharmacol. Sci.* **2008**, *29*, 62–71. [CrossRef]
49. Costin, J.; Jenwitheesuk, E.; Lok, S.; Hunsperger, E.; Conrads, K.; Fontaine, K.; Rees, C.; Rossmann, M.; Isern, S.; Samudrala, R.; et al. Structural optimization and de novo design of dengue virus entry inhibitory peptides. *PLoS Negl. Trop. Dis.* **2010**, *4*, e721. [CrossRef]
50. Palanikumar, L.; Karpauskaite, L.; Al-Sayegh, M.; Chehade, I.; Alam, M.; Hassan, S.; Maity, D.; Ali, L.; Kalmouni, M.; Hunashal, Y.; et al. Protein mimetic amyloid inhibitor potently abrogates cancer-associated mutant p53 aggregation and restores tumor suppressor function. *Nat. Commun.* **2021**, *12*, 3962. [CrossRef] [PubMed]
51. MedicineNet. Available online: https://www.medicinenet.com/me-too_drug/definition.htm (accessed on 30 July 2021).
52. Cheng, G.; Qian, B.; Samudrala, R.; Baker, D. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Res.* **2005**, *33*, 5861–5867. [CrossRef]
53. Xiao, Z.; Morris-Natschke, S.L.; Lee, K.H. Strategies for the optimization of natural leads to anticancer drugs or drug candidates. *Med. Res. Rev.* **2016**, *36*, 32–91. [CrossRef] [PubMed]
54. Sterling, T.; Irwin, J. ZINC 15—Ligand discovery for everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337. [CrossRef]
55. van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
56. Coley, C.; Rogers, L.; Green, W.; Jensen, K. SCScore: Synthetic complexity Learned from a reaction corpus. *J. Chem. Inf. Model.* **2018**, *58*, 251–261. [CrossRef] [PubMed]
57. Bickerton, G.R.; Paolini, G.V.; Besnard, J.; Muresan, S.; Hopkins, A.L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012**, *4*, 90–98. [CrossRef] [PubMed]
58. Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*, 8. [CrossRef]
59. Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: A fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminform.* **2020**, *12*, 70. [CrossRef] [PubMed]
60. Brown, N.; Fiscato, M.; Segler, M.H.; Vaucher, A.C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019**, *59*, 1096–1108. [CrossRef]
61. Li, J.; Zhang, C.X.; Liu, Y.M.; Chen, K.L.; Chen, G. A comparative study of anti-aging properties and mechanism: Resveratrol and caloric restriction. *Oncotarget* **2017**, *8*, 65717–65729. [CrossRef] [PubMed]
62. Soukas, A.; Hao, H.; Wu, L. Metformin as anti-Aging therapy: Is it for everyone? *Trends Endocrinol. Metab.* **2019**, *30*, 745–755. [CrossRef]
63. Aman, Y.; Qiu, Y.; Tao, J.; Fang, E. Therapeutic potential of boosting NAD+ in aging and age-related diseases. *Transl. Med. Aging* **2018**, *2*, 30–37. [CrossRef]
64. Hernández-Camacho, J.; Bernier, M.; López-Lluch, G.; Navas, P. Coenzyme Q10 supplementation in aging and disease. *Front. Physiol.* **2018**, *9*, 44. [CrossRef]
65. Fagot, D.; Pham, D.; Laboureau, J.; Planel, E.; Guerin, L.; Nègre, C.; Donovan, M.; Bernard, B. Crocin, a natural molecule with potentially beneficial effects against skin ageing. *Int. J. Cosmet. Sci.* **2018**, *40*, 388–400. [CrossRef] [PubMed]

66. Bielak-Zmijewska, A.; Grabowska, W.; Ciolko, A.; Bojko, A.; Mosieniak, G.; Bijoch, L.; Sikora, E. The role of curcumin in the modulation of ageing. *Int. J. Mol. Sci.* **2019**, *20*, 1239. [CrossRef]
67. Ichida, J.K.; Blanchard, J.; Lam, K.; Son, E.Y.; Chung, J.E.; Egli, D.; Loh, K.; Carter, A.C.; Di Giorgio, F.P.; Koszka, K.; et al. A small-molecule inhibitor of TGF-β signaling replaces sox2 in reprogramming by inducing nanog. *Cell Stem Cell* **2009**, *5*, 491–503. [CrossRef] [PubMed]
68. Yamoah, E.N.; Li, M.; Shah, A.; Elliott, K.L.; Cheah, K.; Xu, P.X.; Phillips, S.; Young, S.M., Jr.; Eberl, D.F.; Fritzsch, B. Using sox2 to alleviate the hallmarks of age-related hearing loss. *Ageing Res. Rev.* **2020**, *59*, 101042. [CrossRef]
69. Tominaga, K.; Suzuki, H. TGF-β signaling in cellular senescence and aging-related pathology. *Int. J. Mol. Sci.* **2019**, *20*, 5002. [CrossRef] [PubMed]
70. Kennedy, B.K.; Pennypacker, J.K. Drugs that modulate aging: The promising yet difficult path ahead. *Transl. Res.* **2014**, *163*, 456–465. [CrossRef]
71. Jiao, D.; Yang, S. Overcoming resistance to drugs targeting KRASG12C mutation. *Innovation* **2020**, *1*, 100035.
72. Shah, R.; Lester, J. Tyrosine Kinase inhibitors for the treatment of EGFR mutation-positive non-small-cell lung cancer: A clash of the generations. *Clin. Lung Cancer* **2020**, *21*, 216–228. [CrossRef] [PubMed]
73. Wu, Y.L.; Tsuboi, M.; He, J.; John, T.; Grohe, C.; Majem, M.; Goldman, J.W.; Laktionov, K.; Kim, S.W.; Kato, T.; et al. Osimertinib in resected EGFR-mutated Non–Small-Cell Lung Cancer. *N. Engl. J. Med.* **2020**, *383*, 1711–1723. [CrossRef]
74. Molina-Arcas, M.; Moore, C.; Rana, S.; van Maldegem, F.; Mugarza, E.; Romero-Clavijo, P.; Herbert, E.; Horswell, S.; Li, L.S.; Janes, M.R.; et al. Development of combination therapies to maximize the impact of KRAS-G12C inhibitors in lung cancer. *Sci. Transl. Med.* **2019**, *11*, eaaw7999. [CrossRef]
75. Moore, A.; Rosenberg, S.; McCormick, F.; Malek, S. RAS-targeted therapies: Is the undruggable drugged? *Nat. Rev. Drug Discov.* **2020**, *19*, 533–552. [CrossRef]
76. Lau, S.C.M.; Batra, U.; Mok, T.S.K.; Loong, H.H. Dacomitinib in the management of advanced Non-Small-Cell Lung Cancer. *Drugs* **2019**, *79*, 823–831. [CrossRef]
77. Keating, G.M. Afatinib: A review in advanced Non-Small Cell Lung Cancer. *Target. Oncol.* **2016**, *11*, 825–835. [CrossRef] [PubMed]
78. Piperdi, B.; Perez-Soler, R. Role of Erlotinib in the treatment of Non-Small Cell Lung Cancer. *Drugs* **2012**, *72*, 11–19. [CrossRef]
79. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph Neural Networks: A review of methods and applications. *arXiv* **2021**, arXiv:1812.08434.
80. Lamb, J.; Crawford, E.; Peck, D.; Modell, J.; Blat, I.; Wrobel, M.; Lerner, J.; Brunet, J.; Subramanian, A.; Ross, K.; et al. The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* **2006**, *313*, 1929–1935. [CrossRef]
81. Subramanian, A.; Narayan, R.; Corsello, S.; Peck, D.; Natoli, T.; Lu, X.; Gould, J.; Davis, J.; Tubelli, A.; Asiedu, J.; et al. A next generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. *Cell* **2017**, *171*, 1437–1452. [CrossRef] [PubMed]
82. Wleklinski, M.; Loren, B.; Ferreira, C.; Jaman, Z.; Avramova, L.; Sobreira, T.; Thompson, D.; Cooks, R. High throughput reaction screening using desorption electrospray ionization mass spectrometry. *Sci. Rep.* **2018**, *9*, 1647–1653. [CrossRef] [PubMed]
83. Morato, N.; Holden, D.; Cooks, R. High-throughput label-free enzymatic assays using desorption electrospray-ionization mass spectrometry. *Angew. Chem. Int. Ed. Engl.* **2020**, *59*, 20459–20464. [CrossRef]
84. Logsdon, D.; Li, Y.; Sobreira, T.; Ferreira, C.; Thompson, D.; Cooks, R. High-throughput screening of reductive amination reactions using desorption electrospray ionization mass spectrometry. *Org. Process Res. Dev.* **2020**, *24*, 1647–1657. [CrossRef]
85. Sobreira, T.; Avramova, L.; Szilagyi, B.; Logsdon, D.; Loren, B.; Jaman, Z.; Hilger, R.; Hosler, R.; Ferreira, C.; Koswara, A.; et al. High-throughput screening of organic reactions in microdroplets using desorption electrospray ionization mass spectrometry (DESI-MS): Hardware and software implementation. *Methods* **2020**, *12*, 3654–3669. [CrossRef]
86. Le, M.; Morato, N.; Kaerner, A.; Welch, C.; Cooks, R. Fragmentation of polyfunctional compounds recorded using automated high-throughput desorption electrospray ionization. *J. Am. Soc. Mass Spectrom.* **2021**, *32*, 2261–2273. [CrossRef] [PubMed]
87. Morato, N.; Le, M.; Holden, D.; Cooks, R. Automated high-throughput system combining small-scale synthesis with bioassays and reaction screening. *SLAS Technol.* **2021**, *26*, 555–571. [CrossRef]
88. Biyani, S.; Qi, Q.; Wu, J.; Moriuchi, Y.; Larocque, E.A.; Sintim, H.; Thompson, D. Use of high-throughput tools for telescoped continuous flow synthesis of an alkynylnaphthyridine anticancer agent, HSN608. *Org. Process Res. Dev.* **2020**, *24*, 2240–2251. [CrossRef]
89. Jaman, Z.; Mufti, A.; Sah, S.; Avramova, L.; Thompson, D. High Throughput Experimentation and Continuous Flow Validation of Suzuki-Miyaura Cross-Coupling Reactions. *Chem. Eur. J.* **2018**, *24*, 9546–9554. [CrossRef] [PubMed]
90. Wei, Z.; Xie, Z.; Kuvelkar, R.; Shah, V.; Bateman, K.; McLaren, D.; Cooks, R. high-throughput bioassays using "dip-and-go" multiplexed electrospray mass spectrometry. *Angew. Chem. Int. Ed. Engl.* **2019**, *58*, 17594–17598. doi:10.1002/anie.201909047. [CrossRef]
91. Berman, H.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.; Weissig, H.; Shindyalov, I.; Bourne, P. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef] [PubMed]
92. Yang, J.; Zhang, Y. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Res.* **2015**, *43*, W174–W181. [CrossRef]

93. Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: Protein structure and function prediction. *Nat. Methods* **2015**, *12*, 7–8. [CrossRef] [PubMed]
94. Zhang, C.; Mortuza, S.; He, B.; Wang, Y.; Zhang, Y. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins Struct. Funct. Genet.* **2018**, *86* (Suppl. S1), 136–151. [CrossRef]
95. Zhang, W.; Yang, J.; He, B.; Walker, S.; Zhang, H.; Govindarajoo, B.; Virtanen, J.; Xue, Z.; Shen, H.; Zhang, Y. Integration of QUARK and I-TASSER for ab initio protein structure prediction in CASP11. *Proteins Struct. Funct. Genet.* **2016**, *84* (Suppl. S1), 76–86. [CrossRef]
96. Yang, J.; Zhang, W.; He, B.; Walker, S.; Zhang, H.; Govindarajoo, B.; Virtanen, J.; Xue, Z.; Shen, H.; Zhang, Y. Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade. *Proteins Struct. Funct. Genet.* **2016**, *84* (Suppl. S1), 233–246. [CrossRef]
97. Xu, D.; Zhang, J.; Roy, A.; Zhang, Y. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins Struct. Funct. Genet.* **2011**, *79* (Suppl. S10), 147–160. [CrossRef] [PubMed]
98. Wu, Q.; Peng, Z.; Zhang, Y.; Yang, J. COACH-D: Improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. *Nucleic Acids Res.* **2018**, *46*, W438–W430. [CrossRef] [PubMed]
99. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **2006**, *11*, 1046–1053. [CrossRef] [PubMed]
100. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inform. Comput. Sci.* **1988**, *28*, 31–36. [CrossRef]
101. Kramer, M. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J.* **1991**, *37*, 233–243. [CrossRef]
102. Santhanam, S. Context based text-generation using LSTM networks. *arXiv* **2020**, arXiv:2005.00048.
103. Hancock, J.T.; Khoshgoftaar, T.M. Survey on categorical data for neural networks. *J. Big Data* **2020**, *7*, 28. [CrossRef]
104. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
105. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A Review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef] [PubMed]
106. Kullback, S.; Leibler, R. On information and sufficiency. *Ann. Math. Statist.* **1951**, *22*, 79–86. [CrossRef]
107. Davis, A.; Grondin, C.; Johnson, R.; Sciaky, D.; Wiegers, J.; Wiegers, T.; Mattingly, C. Comparative Toxicogenomics Database (CTD): Update 2021. *Nucleic Acids Res.* **2021**, *49*, D1138. [CrossRef]
108. Tatonetti, N.; Ye, P.; Daneshjou, R.; Altman, R. Data-driven prediction of drug effects and interactions. *Sci. Transl. Med.* **2012**, *4*, 125–129. [CrossRef]
109. Kuhn, M.; Letunic, I.; Jensen, L.; Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **2016**, *44*, D1075–D1079. [CrossRef]
110. Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **2015**, *7*, 20. [CrossRef] [PubMed]
111. Feller, W. On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions. *Ann. Math. Stat.* **1948**, *19*, 177–189. [CrossRef]
112. Karson, M. Handbook of Methods of Applied Statistics. Volume I: Techniques of Computation Descriptive Methods, and Statistical Inference. Volume II: Planning of Surveys and Experiments. I. M. Chakravarti, R. G. Laha, and J. Roy, New York, John Wiley; 1967, $9.00. *J. Am. Stat. Assoc.* **1968**, *63*, 1047–1049. [CrossRef]
113. Cox, D.R. Statistical significance tests. *Br. J. Clin. Pharmacol.* **1982**, *14*, 325–331. [CrossRef]
114. Johnson, A.A.; Shokhirev, M.N.; Wyss-Coray, T.; Lehallier, B. Systematic review and analysis of human proteomics aging studies unveils a novel proteomic aging clock and identifies key processes that change with age. *Ageing Res. Rev.* **2020**, *60*, 101070. [CrossRef]
115. Lorusso, J.S.; Sviderskiy, O.A.; Labunskyy, V.M. Emerging omics approaches in aging research. *Antioxid. Redox Signal.* **2018**, *29*, 985–1002. [CrossRef] [PubMed]
116. Gill, D.; Parry, A.; Santos, F.; Hernando-Herraez, I.; Stubbs, T.M.; Milagre, I.; Reik, W. Multi-omic rejuvenation of human cells by maturation phase transient reprogramming. *bioRxiv* **2021**. . [CrossRef]
117. Natarajan, K.N.; Teichmann, S.A.; Kolodziejczyk, A.A. Single cell transcriptomics of pluripotent stem cells: reprogramming and differentiation. *Curr. Opin. Genet. Dev.* **2017**, *46*, 66–76. [CrossRef] [PubMed]
118. Ramsay, R.R.; Popovic-Nikolic, M.R.; Nikolic, K.; Uliassi, E.; Bolognesi, M.L. A perspective on multi-target drug discovery and design for complex diseases. *Clin. Transl. Med.* **2018**, *7*, 3. [CrossRef]
119. Koomen, J.M.; Haura, E.B.; Bepler, G.; Sutphen, R.; Remily-Wood, E.R.; Benson, K.; Hussein, M.; Hazlehurst, L.A.; Yeatman, T.J.; Hildreth, L.T.; et al. Proteomic contributions to personalized cancer care. *Mol. Cell. Proteom.* **2008**, *7*, 1780–1794. [CrossRef] [PubMed]
120. Yumura, M.; Nagano, T.; Nishimura, Y. Novel multitarget therapies for lung cancer and respiratory disease. *Molecules* **2020**, *25*, 3987. [CrossRef] [PubMed]