International Journal of
Molecular Sciences

MDPI

Article

TumFlow: An AI Model for Predicting New Anticancer Molecules

Davide Rigoni [1,*], Sachithra Yaddehige [2], Nicoletta Bianchi [3], Alessandro Sperduti [4], Stefano Moro [1] and Cristian Taccioli [2]

[1] Molecular Modelling Section (MMS), Department of Pharmaceutical and Pharmacological Sciences, University of Padova, Via Francesco Marzolo 5, 35131 Padova, Italy; stefano.moro@unipd.it

[2] Department of Animal Medicine, Production and Health, University of Padova, Viale dell'Università 16, 35020 Legnaro, Italy; sachithrakalhari.yaddehige@studenti.unipd.it (S.Y.); cristian.taccioli@unipd.it (C.T.)

[3] Department of Translational Medicine, University of Ferrara, Via Luigi Borsari 46, 44121 Ferrara, Italy; nicoletta.bianchi@unife.it

[4] Department of Mathematics "Tullio Levi-Civita", University of Padova, Via Trieste 63, 35131 Padova, Italy; alessandro.sperduti@unipd.it

* Correspondence: davide.rigoni.1@unipd.it

Abstract: Melanoma is the fifth most common cancer in the United States. Conventional drug discovery methods are inherently time-consuming and costly, which imposes significant limitations. However, the advent of Artificial Intelligence (AI) has opened up new possibilities for simulating and evaluating numerous drug candidates, thereby mitigating the requisite time and resources. In this context, normalizing flow models by employing machine learning techniques to create new molecular structures holds promise for accelerating the discovery of effective anticancer therapies. This manuscript introduces *TumFlow*, a novel AI model designed to generate new molecular entities with potential therapeutic value in cancer treatment. It has been trained on the NCI-60 dataset, encompassing thousands of molecules tested across 60 tumour cell lines, with an emphasis on the melanoma SK-MEL-28 cell line. The model successfully generated new molecules with predicted improved efficacy in inhibiting tumour growth while being synthetically feasible. This represents a significant advancement over conventional generative models, which often produce molecules that are challenging or impossible to synthesize. Furthermore, *TumFlow* has also been utilized to optimize molecules known for their efficacy in clinical melanoma treatments. This led to the creation of novel molecules with a predicted enhanced likelihood of effectiveness against melanoma, currently undocumented on PubChem.

Keywords: generative model; anticancer molecules; melanoma; SK-MEL-28

check for updates

Citation: Rigoni, D.; Yaddehige, S.; Bianchi, N.; Sperduti, A.; Moro, S.; Taccioli, C. TumFlow: An AI Model for Predicting New Anticancer Molecules. *Int. J. Mol. Sci.* **2024**, *25*, 6186. https://doi.org/10.3390/ijms25116186

Academic Editor: Renata Tisi

Received: 22 April 2024
Revised: 22 May 2024
Accepted: 23 May 2024
Published: 4 June 2024

1. Introduction

Melanoma, a serious form of skin cancer, originates from melanocytes. Melanocytes are cells responsible for producing melanin that colours the skin. It stands as the most severe type of skin cancer due to its potential to metastasize to other body parts if not detected and treated promptly. Individuals with fair skin, blue eyes, and light-coloured hair are predominantly at higher risk, largely due to their lower levels of melanin, making their skin more susceptible to harmful ultraviolet (UV) radiation from the sun [1–5]. Moreover, melanoma poses an increased threat due to its resistance to conventional chemotherapy [6]. Current treatment strategies for melanoma include surgical excision, targeted therapy, and immunotherapy. Targeted therapies are employed for melanomas with specific genetic mutations, such as the BRAF V600E mutation, using inhibitors like vemurafenib and dabrafenib [7]. Immunotherapy, leveraging agents such as anti-PD-1 antibodies (nivolumab and pembrolizumab) and anti-CTLA-4 antibodies (ipilimumab), has shown efficacy in enhancing the immune response against melanoma cancer cells [8–10]. The drug discovery and design processes are complex and resource-intensive, often extending over 10–20 years with costs exceeding USD 2 billion [11,12]. Figure S1 presents the number of FDA-approved drugs per year [13], highlighting the small increase in approvals despite investments in research and development increasing each year [14], as can be seen in Figure S2.

In this context, Artificial Intelligence (AI) provides a promising avenue for revolutionizing the field, potentially reducing costs and increasing efficiency. It has become a pivotal tool in various aspects of cancer management, encompassing early detection, precision medicine, imaging, and drug repurposing [15]. Despite these advancements, the complete potential of AI in synthesizing novel anticancer molecules is yet to be fully harnessed and explored [16]. Within AI, machine learning and deep learning are key subfields, including techniques like supervised and unsupervised learning. Supervised learning is utilized for tasks like disease detection and drug efficiency estimation, while unsupervised learning aids in patient stratification and disease recognition [17]. Deep learning, particularly effective in processing large datasets such as those related to the use of images, has contributed notably to melanoma cancer diagnostics among other areas [18–20].

Among the unsupervised models, there is a family of approaches that fall under the name of generative models, which are a class of algorithms designed to learn and generate new data that are similar to those within a given training dataset. These models aim to capture the underlying patterns and structures in the training data, enabling them to generate novel samples that share characteristics with the original data. In the field of new drug generation, various approaches based on Variational Autoencoders (VAEs) [21–31], Generative Adversarial Networks (GANs) [32–34], Normalizing Flows [11,35–39], and Diffusion Models [40–50] have been explored. Moreover, the advent of large language models (LLMs) using transformer architectures [51] has further expanded this field. Transformer-based models, originally developed for natural language processing tasks, have been successful in capturing complex patterns in data and have been applied in drug generation [52–55]. Generative models can also include a predictive model to predict the antitumoral activity of generated molecules, enabling the identification of the most promising candidates.

Normalizing flow methods have been applied in various fields, including density estimation [56] and data augmentation [57]. They have been used in tasks such as image generation, speech synthesis, and molecular generation in chemistry. Specific architectures like Real Non-Volume-Preserving (RealNVP) [58] and Glow [59] methods are examples of these. The normalizing flow method represents an effective technique to learn the unknown probability distribution that has generated the data in the training set, i.e., the chemical structure of the molecules. It does this by employing a series of invertible transformations to transmute a probability distribution over input data (i.e., molecule structures) into a designated target probability distribution.

This research incorporates deep learning into drug discovery with *TumFlow*, a novel approach for generating molecular graphs for cancer therapeutics. *TumFlow*, building on the foundational work of MoFlow [11], a pioneering model in the field of models applied to graph structures, adapts and enhances these capabilities specifically to address melanoma treatment challenges. It leverages MoFlow efficient bond and atom generation to create novel molecules aimed to be effective against melanoma cancer cells. When learning to generate new antitumoral molecules, *TumFlow* is trying to solve a complex assignment made of challenging subtasks. The successful generation of useful molecules requires an implicit comprehension of their pharmacokinetics, the identification of single or multiple targets, and the assurance that they bind with high affinity to these to inhibit tumour progression. Each of these subtasks presents formidable difficulties independently, and the fact that the neural network does not have this kind of information to learn from makes the learning process even more challenging.

The integration of *TumFlow* into the drug discovery process reflects a broader trend in AI increasing impact on healthcare and pharmaceutical research. By focusing specifically on melanoma, *TumFlow* addresses a critical need in cancer treatment, offering the potential to rapidly identify and develop new therapeutic molecules. For this reason, this work represents a novel contribution to anticancer drug discovery.

2. Results and Discussion

In the following, some novel molecules generated by *TumFlow* against the SK-MEL-28 melanoma tumour are presented, while its limitations are discussed in Section S7 of the Supplementary Material. Two processes were adopted for generating new molecules with *TumFlow* that will be individually discussed in the following. Each molecule will be introduced with its predicted GI50 score, the normalized SAS [60] value, and its similarity measure in relation to the initial molecule. The similarity score allows evaluation of how much the newly generated molecule and the starting molecule differ from each other. This score is calculated using the Tanimoto similarity of the Morgan Fingerprint [61]. More details about the generation process adopted by *TumFlow* and the metrics considered in this work are reported in Section 3.

2.1. Generation Starting from the NCI-60 Dataset

This section presents a chosen set of novel molecules generated by *TumFlow* considering molecules from the training set as starting points. Specifically, in this generation procedure, the first 140 molecules with higher antitumoral efficacy appearing in the training set, i.e., antitumoral molecules tested in vitro from the NCI-60 project [62], were used as a starting point.

Figure 1 presents some molecules obtained from a provided starting molecule, while the corresponding canonical SMILES [63–65] are reported in Table S2. Specifically, the figure presents a grid where the first column on the left depicts the starting molecule structures, while the other columns report the new molecules obtained from them. By inspecting the molecular structures and the corresponding predicted GI50 scores, it can be seen that *TumFlow* attempts to enhance the structure of the provided starting molecule to improve its efficacy against melanoma tumours. Nevertheless, in the process, the model tends to generate increasingly complex molecules, introducing synthesis issues, and molecular structures dissimilar from the starting one.

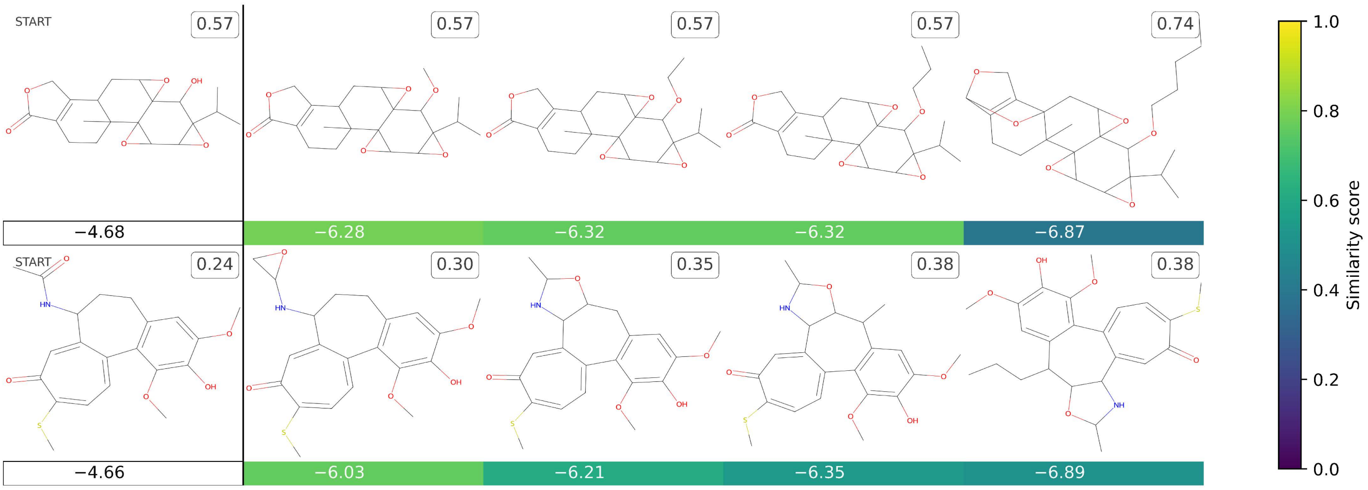

Figure 1. This grid presents the novel molecules that *TumFlow* generated starting from those in the dataset. The first column on the left reports the starting molecule structures, while the other columns report the new molecules. The score reported under each generated molecule represents the *TumFlow* predicted GI50 score, while the colour conveys the similarity score of the newly generated structure in relation to the starting molecule structure. The number to the top right of each molecule is the normalized SAS score.

Contrary to Figure 1, Figure 2 presents a few molecules obtained by the whole generation process, i.e., considering many starting structures. The corresponding canonical SMILES, including those of the provided starting molecules, are reported in Table S3. All of these molecules are chemically noteworthy and interesting, particularly due to their absence in the dataset. It is important to note that the generation process sometimes results in uncommon molecules that encounter challenges in synthesis and/or contain rare substructures. Nevertheless, thanks to the SAS score, it becomes possible to identify molecules that are challenging to synthesize and, consequently, filter them as necessary.

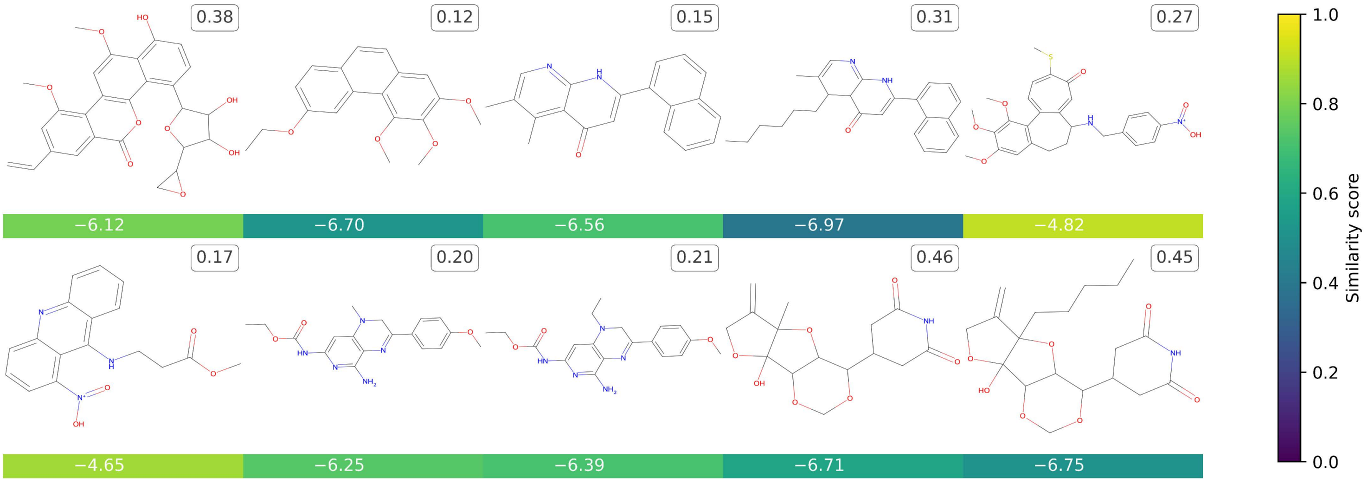

Figure 2. This grid presents the novel molecules that *TumFlow* generated starting from those in the dataset. The score reported under each generated molecule represents the *TumFlow* predicted GI50 score, while the colour conveys the similarity score of the newly generated structure in relation to the starting molecule structure. The number to the top right of each molecule is the normalized SAS score.

None of the novel molecules, except the first compound (CID = 121297650) generated by the first molecule reported in Figure 1, are available in PubChem [66]. This absence of novel molecules in PubChem highlights the pioneering nature of *TumFlow* in exploring unknown chemical spaces, while the presence of the already-existing molecule demonstrates the ability of the model to generate meaningful structures.

2.2. Generation Starting from Clinically Adopted Anti-Melanoma Molecules

Herein, a selected set of novel molecules generated by *TumFlow* considering clinical molecules as starting points are presented. Specifically, some of the molecules reported in Table S1, known for their efficacy in clinical treatments for melanoma, were used as starting points.

Figure 3 presents some novel molecules, while the corresponding canonical SMILES are reported in Table S4. Regarding the molecules generated from clinical drugs, a pattern akin to those originating from in vitro molecules is discernible. In this scenario as well, *TumFlow* demonstrates the capacity to generate novel molecular structures, albeit occasionally encountering challenges in synthesis. Notably, with the exception of just one molecule, all the newly generated structures are absent from PubChem. In fact, the initial molecule derived from the first clinical drug, specifically the second structure in the first row of the image, has been identified as a previously studied compound against cancer, bearing the corresponding NSC = 133726 and CID = 421441. More precisely, this compound has already been subjected to several **in vivo testing** on mice, demonstrating its activity against the leukemia cell line L1210 (e.g., PubChem AID = 248).

The identification of a novel molecule, previously studied for its anticancer properties and **not present in the training set**, underscores *TumFlow*'s potential ability to explore the chemical space beyond the confines of existing datasets. This capability suggests that *TumFlow* has the capacity to propose compounds with therapeutic relevance that might not have been part of the original training data, stressing its potential to contribute to the discovery of compounds with valuable properties.

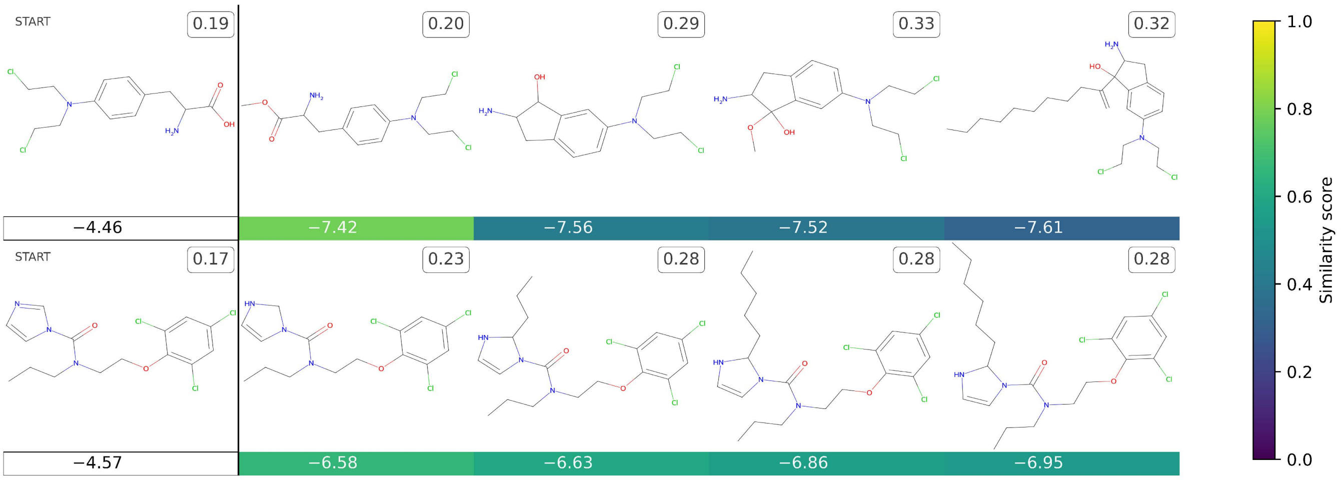

Figure 3. This grid presents the novel molecules that *TumFlow* generated starting from two clinical molecules. The first column reports the starting molecule structures, while the other columns report the new molecules. The score reported under each generated molecule represents the *TumFlow* predicted GI50 score, while the colour conveys the similarity score of the newly generated structure in relation to the starting molecule structure. The number to the top right of each molecule is the normalized SAS score.

2.3. Benchmarking

During the development and evaluation of *TumFlow*, we encountered significant challenges: the absence of a benchmarking dataset suitable for model validation and the intrinsic problem of synthesizing new molecules. Benchmark datasets play a crucial role in the field of machine learning, providing a base against which newly developed models can be tested and compared. These datasets enable researchers to assess the accuracy, efficiency, and overall performance of their models in a standardized context. Unfortunately, in the context of generating novel molecular entities for treating the SK-MEL-28 tumour, such a benchmarking dataset does not exist.

The lack of a benchmark could be compensated by experimental validation. Yet, experimentally validating the molecules produced by *TumFlow* also presents its challenges. As *TumFlow* designs new molecular structures with potential anticancer properties, most of the candidate molecules lack prior synthesis or testing documentation. Consequently, before any biological efficacy testing can start (e.g., assays on melanoma cell lines), these molecules must first be synthesized. Synthesizing new molecules is not only a complex process that demands specialized expertise but also involves substantial time investment and significant financial costs.

These challenges present a substantial hurdle for validating the effectiveness of generative approaches in generating therapeutically valuable molecules.

2.4. Code Implementation

The implementation of the *TumFlow* model, along with the code utilized for training and generating novel molecules and a comprehensive user configuration guide, are openly available at https://github.com/drigoni/TumFlow (accessed on 6 February 2024). Within the repository, there are all the necessary scripts for result reproducibility, enabling robust verification of findings. Furthermore, the repository hosts the trained weights of the *TumFlow* model as well as the dataset, including all GI50, IC50, LC50 and TGI scores used in this work. Additionally, a Docker [67] container is provided to streamline usage across various computing environments, ensuring accessibility and ease of deployment.

3. Materials and Methods

TumFlow is based on MoFlow, a normalizing flow model originally developed for the generation of graph structures without any focus on anticancer molecules. On the other hand, *TumFlow* aims to learn the unknown probability distribution that has generated the chemical structures of the molecules in the dataset, with the purpose of using the learned distribution to generate new novel chemical structures that should convey similar substructures and similar anticancer activities. Therefore, *TumFlow* is developed to predict new antitumour molecules against the SK-MEL-28 melanoma, addressing all the unique challenges and requirements of melanoma treatment. It is trained on the comprehensive NCI-60 dataset, made public by the National Cancer Institute [68], which encompasses thousands of molecules tested across a broad spectrum of tumour cell lines.

The following sections will present in more detail the data preprocessing method applied to the NCI-60 dataset, as well as the *TumFlow* model. Additional details about normalizing flows are reported in Section S3 of the Supplementary Material, while more details on the *TumFlow* model are reported in Section S4.

The following mathematical notations are adopted:

(i) Lower-case symbols for scalars, indexes, and assignment to random variables, e.g., n and x;
(ii) Italic upper-case symbols for sets and single random variables, e.g., A and X;
(iii) Bold lower-case symbols for vectors and assignments to vectors of random variables, e.g., $\boldsymbol{a}$ and $\boldsymbol{x}$;
(iv) Bold upper-case symbols for matrices, tensors, and vectors of random variables, e.g., $\boldsymbol{A}$ and $\boldsymbol{Z}$;
(v) The position within a tensor or vector is denoted by numeric subscripts in square brackets, for example, $\boldsymbol{A}_{[i,a:b,:]}$, where $i, a, b \in \mathbb{N}^+$, and ":" indicates the positions from a to b. The solitary use of the colon symbol ":" represents all positions;
(vi) Calligraphic symbols for domains, e.g., $\mathcal{Q}$;
(vii) When it is clear from the context, the probability random variables are omitted, as $\mathbb{P}(x)$ instead of $\mathbb{P}(X = x)$.

3.1. Data Sources and Data Preprocessing

The NCI-60 project [62], launched in 1990, employs 60 human tumour cell lines representing diverse cancers to evaluate up to 7000 small molecules annually for anticancer properties. It provides four files with the results of their experiments: "GI50.csv", "LC50.csv", "IC50.csv", and "TGI.csv". In this work, only the "GI50.csv" file was used. It contains data on the GI50 (Growth Inhibition of 50%) values, which are derived from laboratory assays that measured the concentration of a chemical compound required to inhibit the growth of a specific tumour cell line by 50%. Specifically, the GI50 values are obtained by interpolating the *GIPRCNT* scores, which are the percentage of treated cell growth as a fraction of control cell growth, corrected for the count of cells at the time of drug addition in the assay. A score of 100 is control growth, 0 is complete inhibition of growth (cytostasis), and -100 is complete cell kill. Thus, these values serve as a direct indicator of the compound's potential antitumoral efficacy as a lower GI50 value indicates a higher efficacy of the molecule in inhibiting tumour growth in the tested cells. More information is reported in the NCI-60 project website. In addition, the dataset includes the National Service Center (NSC) code, a unique numeric identifier assigned to substances tested and evaluated by the National Cancer Institute, and information on the tested cell line.

The choice of this file was based on its relevance in identifying compounds with potential antitumoral efficacy, particularly in the context of SK-MEL-28 melanoma cells. In fact, a preliminary data analysis visible in Figure S3 revealed that molecules used clinically show better representation in the GI50 dataset. Indeed, the GI50 scores offer a more accurate representation of clinical drugs since they exhibit a more evenly distributed pattern and better distinguish the effects of various drugs. Additionally, the violin plot illustrating the mass of GI50 scores demonstrates greater variability than that of the IC50 scores, which, conversely, appear more condensed. However, even though the presented work focuses on the GI50 score, other indicators like the IC50 can also be utilized seamlessly. This correlation reinforces the validity of the approach presented in this work and highlights the importance of integrating real and clinically relevant data into the modelling process.

The training of the *TumFlow* model was performed on this data, focusing only on molecules tested on SK-MEL-28 melanoma cell lines made by chemical elements commonly found in organic compounds (only molecules composed by hydrogen (H), carbon (C), nitrogen (N), oxygen (O), fluorine (F), phosphorus (P), sulphur (S), chlorine (Cl), selenium (Se), bromine (Br), and iodine (I)). During the data preprocessing phase, all molecules with a positively charged oxygen (O^+) were removed, and all "ion pair" compounds were sanitized, selecting only the largest connected component as the main molecule structure while discarding the remaining smaller component(s). If the sanitization resulted in a structure already existing in the dataset, only the experiment with the highest efficacy score was retained. For molecules with multiple in vitro experiments, the corresponding GI50 values were averaged. Following the data preprocessing phase, the dataset consists of 46,766 unique molecules, each paired with its corresponding GI50 efficacy value.

3.2. TumFlow

TumFlow aims to predict new anticancer molecules by exploiting the graph representation of the molecule structure, differently from other works adopting the SMILES sequential representation of the molecule, such as [22,23].

Mathematically, let $\mathcal{D}$ be the dataset, Tr be the training set of molecules, Θ_n and Θ_e be, respectively, the set of atom types and the set of edge types extracted from dataset $\mathcal{D}$. Let $d_v = \mid \Theta_n \mid$ be the number of atom types, $d_e = \mid \Theta_e \mid$ be the number of edge types, and d_n the maximum number of atoms, hydrogens excluded, forming the molecules in dataset $\mathcal{D}$. Then, a molecule is represented as a graph $G \in \mathcal{G}$:

$$G = (\boldsymbol{V}, \boldsymbol{E}),$$

where $\boldsymbol{V} \in \{0,1\}^{d_n \times d_v}$ is a node-type matrix and $\boldsymbol{E} \in \{0,1\}^{d_n \times d_n \times d_e}$ is an edge-type tensor, such that $\boldsymbol{V}_{i,v} = 1$ only if the molecule node i is of type v and such that $\boldsymbol{E}_{i,j,e} = 1$ only if the molecule nodes i and j are connected through a bond of type e. The set of all possible graphs is defined as follows:

$$\mathcal{G} = \mathcal{V} \times \mathcal{E} = \{0,1\}^{d_n \times d_v} \times \{0,1\}^{d_n \times d_n \times d_e}.$$

TumFlow aims to learn the complex probability distribution $\mathbb{P}_{\mathcal{G}}$, from which the molecules in the dataset are generated, in order to sample from it new useful molecule graphs $G \sim \mathbb{P}_{\mathcal{G}}\left(\widetilde{G}\right)$. $\widetilde{G}$ denotes the random variable over graph structures with support in $\mathcal{G}$. *TumFlow* factorizes the probability distribution as follows:

$$\mathbb{P}_{\mathcal{G}}\left(\widetilde{G}\right) = \mathbb{P}_{\mathcal{G}}\left(\left(\widetilde{\boldsymbol{V}}, \widetilde{\boldsymbol{E}}\right)\right) = \mathbb{P}_{\mathcal{V}}\left(\widetilde{\mathbf{V}} \mid \widetilde{\boldsymbol{E}}\right) \cdot \mathbb{P}_{\mathcal{E}}\left(\widetilde{\boldsymbol{E}}\right),$$

where $\mathbb{P}_{\mathcal{E}}\left(\widetilde{\boldsymbol{E}}\right)$ is the probability distribution over molecule bounds, $\mathbb{P}_{\mathcal{V}}\left(\widetilde{\boldsymbol{V}} \mid \widetilde{\boldsymbol{E}}\right)$ is the conditioned probability distribution over atoms given molecule bounds, and both $\widetilde{\boldsymbol{E}}$ and $\widetilde{\boldsymbol{V}}$ are vectors of random variables. In simpler terms, *TumFlow* first predicts the set of bonds forming the structure of the molecule and then conditions the generation of the molecule atoms by the predicted bonds. It uses two jointly trained normalizing flow models. The first is used to predict the molecule bonds and the second is used to predict the molecule atoms. The decision to factorize the full probability distribution as predicting the bonds forming the molecule's structure before predicting the atoms is purposeful. This factorization enables the effective utilization of graph neural networks (more on this in the subsequent paragraphs). In graph neural networks, nodes update their states based on information propagated through the edges. Thus, by first predicting the bonds (edges), a foundation is established upon which the subsequent prediction of atoms (nodes) can be informed and influenced. This approach aligns well with the nature of molecular structures, where the connectivity between atoms greatly influences their properties and behaviours. Figure 4 reports the overall model architecture, summarizing the main steps.

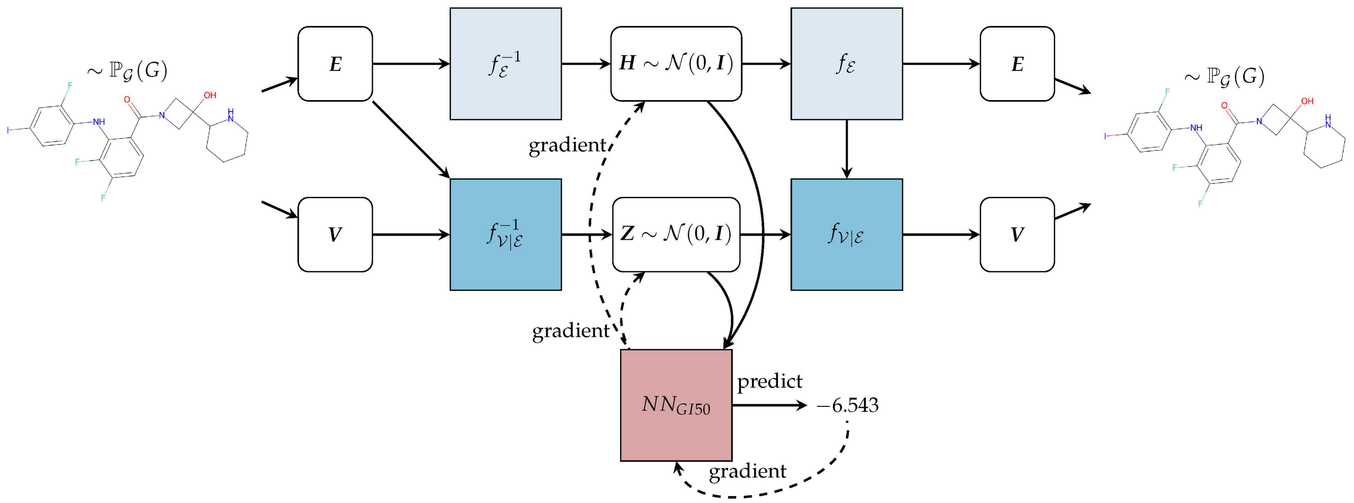
$\sim \mathbb{P}_{\mathcal{G}}(G)$
F
NH
O
OH
N
H
I
$\boldsymbol{E}$
$f_{\mathcal{E}}^{-1}$
$\boldsymbol{H} \sim \mathcal{N}(0, \boldsymbol{I})$
$f_{\mathcal{E}}$
$\boldsymbol{E}$
$\sim \mathbb{P}_{\mathcal{G}}(G)$
gradient
$\boldsymbol{V}$
$f_{\mathcal{V}|\mathcal{E}}^{-1}$
$\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{I})$
$f_{\mathcal{V}|\mathcal{E}}$
$\boldsymbol{V}$
gradient
NN_{GI50}
predict
−6.543
gradient

Figure 4. Overview of the *TumFlow* model. From the molecule in input (on the left) the graph $G = (\boldsymbol{V}, \boldsymbol{E})$ is constructed, and both the latent representations $\boldsymbol{H} = f_{\mathcal{E}}^{-1}(\boldsymbol{E})$ and $\boldsymbol{Z} = f_{\mathcal{V}|\mathcal{E}}^{-1}(\boldsymbol{V}; \boldsymbol{E})$ are obtained. From the latent representations, the graph G and then the molecule in output (on the right) are reconstructed through the functions $f_{\mathcal{E}}$ and $f_{\mathcal{V}|\mathcal{E}}$. The module NN_{GI50} predicts the GI50 score and can optimize the molecule structure, employing the gradient descent approach.

TumFlow is trained to optimize the negative log-likelihood loss:

$$\begin{aligned}\mathcal{L}_{\mathcal{G}}(\mathcal{D}) &= -\frac{1}{\mid \mathcal{D} \mid} \sum_{G \in \mathcal{D}} \log \mathbb{P}_{\mathcal{G}}(G); \\ &= -\frac{1}{\mid \mathcal{D} \mid} \sum_{(\boldsymbol{V}, \boldsymbol{E}) \in \mathcal{D}} \log \mathbb{P}_{\mathcal{V}}(\boldsymbol{V} \mid \boldsymbol{E}) + \log \mathbb{P}_{\mathcal{E}}(\boldsymbol{E});\end{aligned}$$

with

$$\begin{aligned}\log \mathbb{P}_{\mathcal{E}}(\boldsymbol{E}) &= \log \mathbb{P}_{\boldsymbol{H}}(\boldsymbol{H}) - \log \left| \det\left(\frac{\partial f_{\mathcal{E}}(\boldsymbol{H})}{\partial \boldsymbol{H}} \right) \right|; \\ \log \mathbb{P}_{\mathcal{V}}(\boldsymbol{V} \mid \boldsymbol{E}) &= \log \mathbb{P}_{\boldsymbol{Z}}(\boldsymbol{Z} \mid \boldsymbol{E}) - \log \left| \det\left(\frac{\partial f_{\mathcal{V}|\mathcal{E}}(\boldsymbol{Z}; \boldsymbol{E})}{\partial \boldsymbol{Z}} \right) \right|;\end{aligned}$$

where $f_{\mathcal{E}}$ and $f_{\mathcal{V}|\mathcal{E}}$ are two invertible and differentiable functions to learn, $\boldsymbol{H}$ and $\boldsymbol{Z}$ are, respectively, two latent representations for atom and adjacency tensors, and $\mathbb{P}_{\boldsymbol{H}}$ and $\mathbb{P}_{\boldsymbol{Z}}$ are the two simple target distributions, i.e., two standard normal distributions $\mathcal{N}(0, \boldsymbol{I})$, with zero mean and identity matrix as covariance matrix.

Affine coupling layers are used in the implementation of both $f_{\mathcal{V}|\mathcal{E}} = \Phi_{l_{\mathcal{V}|\mathcal{E}}} \circ \ldots \circ \Phi_1$ and $f_{\mathcal{E}} = \Psi_{l_{\mathcal{E}}} \circ \ldots \circ \Psi_1$, where $l_{\mathcal{V}|\mathcal{E}}$ and $l_{\mathcal{E}}$ represent the number of coupling layers composing $f_{\mathcal{V}|\mathcal{E}}$ and $f_{\mathcal{E}}$, respectively. For the sake of clarity, the explicit dependency on $\boldsymbol{E}$ in the notation of Φ_i is omitted. In the implementation of the coupling layers, the sigmoid function replaces the exponential function, as it provides better numerical stability when stacking multiple coupling layers. Mathematically, each function $\Phi_i^{-1} : \mathbb{R}^{d_n \times d_v} \to \mathbb{R}^{d_n \times d_v} \forall i \in \left\{1, \ldots, l_{\mathcal{V}|\mathcal{E}}\right\}$ splits the input into two parts according to the node types dimension d_v. Given $\boldsymbol{Z}^{i-1} = \Phi_i^{-1}\left(\boldsymbol{Z}^i\right)$ and a selected dimension $\tilde{d_v}$:

$$\boldsymbol{Z}^{i-1} = \begin{cases} \boldsymbol{Z}^{i-1}_{[:,1:\tilde{d_v}]} &= \boldsymbol{Z}^{i}_{[:,1:\tilde{d_v}]}; \\ \boldsymbol{Z}^{i-1}_{[:,\tilde{d_v}+1:d_n]} &= \boldsymbol{Z}^{i}_{[:,\tilde{d_v}+1:d_n]} \odot sig\left(s^i_{\mathcal{V}|\mathcal{E}}\left(\boldsymbol{Z}^{i}_{[:,1:\tilde{d_v}]}\right)\right) + t^i_{\mathcal{V}|\mathcal{E}}\left(\boldsymbol{Z}^{i}_{[:,1:\tilde{d_v}]}\right); \end{cases}$$

where $\boldsymbol{Z} = \boldsymbol{Z}^0$ and $\boldsymbol{V} = \boldsymbol{Z}^{l_{\mathcal{V}|\mathcal{E}}}$. Functions $s^i_{\mathcal{V}|\mathcal{E}}$ and $t^i_{\mathcal{V}|\mathcal{E}}$ are multi-layer perceptions (MLPs) based on the output of a graph neural network [69], which aims to learn the representation of the graph underlying the molecular structure.

Similarly, each function $\Psi_i^{-1} : \mathbb{R}^{d_n \times d_n \times d_e} \rightarrow \mathbb{R}^{d_n \times d_n \times d_e} \forall i \in \{1, \ldots, l_{\mathcal{E}}\}$ splits the input into two parts according to the bond types dimension d_e. Given $\boldsymbol{H}^{i-1} = \Psi_i^{-1}\left(\boldsymbol{H}^i\right)$ and a selected dimension $\widetilde{d}_e$,

$$\boldsymbol{H}^{i-1} = \begin{cases} \boldsymbol{H}^{i-1}_{[:,:,1:\widetilde{d}_e]} & = \boldsymbol{H}^{i}_{[:,:,1:\widetilde{d}_e]}\,; \\ \boldsymbol{H}^{i-1}_{[:,:,\widetilde{d}_e+1:d_e]} & = \boldsymbol{H}^{i}_{[:,:,\widetilde{d}_e+1:d_e]} \odot sig\left(s^i_{\mathcal{E}}\left(\boldsymbol{H}^{i}_{[:,:,1:\widetilde{d}_e]}\right)\right) + t^i_{\mathcal{E}}\left(\boldsymbol{H}^{i}_{[:,:,1:\widetilde{d}_e]}\right), \end{cases}$$

where $\boldsymbol{H} = \boldsymbol{H}^0$ and $\boldsymbol{E} = \boldsymbol{H}^{l_{\mathcal{E}}}$. Functions $s^i_{\mathcal{E}}$ and $t^i_{\mathcal{E}}$ are implemented with a sequence of 2D convolutional neural networks.

It is essential to consider that the normalizing flow framework is designed for continuous space values, and, as such, it cannot be directly applied to discrete structures like node and adjacency tensors. To address this limitation, a pre-processing step is implemented before the utilization of coupling layers. Specifically, a random uniform noise drawn from a carefully selected interval of values is added to each entry of the tensors. This introduction of noise serves the purpose of incorporating a continuous element into the discrete structures, aligning them with the framework requirements and enabling the subsequent application of coupling layers. The carefully selected noise distribution enables the accurate selection of corresponding atoms and bonds through the argmax function when utilizing $f_{\mathcal{V}|\mathcal{E}}$ and $f_{\mathcal{E}}$ to reconstruct G.

3.2.1. Prediction of the GI50 Scores

TumFlow includes a nonlinear neural network NN_{GI50} designed to predict the antitumour activity of individual molecules. This neural network is learned once the main normalizing flow networks have been learned. More precisely, the neural network is trained to predict the GI50 score associated with the molecule's latent representation. Mathematically, NN_{GI50} represents the following function:

$$NN_{GI50} : \mathcal{G} \rightarrow \mathbb{R}.$$

This function is trained with the mean squared error (squared L2 norm) loss computed among predicted values and values measured in vitro. In more detail, given a molecule graph G, its predicted antitumour activity p is estimated as follows:

$$\begin{aligned} p &= NN_{GI50}(f_{vc}(\boldsymbol{H}, \boldsymbol{Z})); \\ \boldsymbol{H} &= f_{\mathcal{E}}^{-1}(\boldsymbol{E}); \\ \boldsymbol{Z} &= f_{\mathcal{V}|\mathcal{E}}^{-1}(\boldsymbol{V}; \boldsymbol{E}); \end{aligned}$$

where f_{vc} is a function that linearizes both the tensors $\boldsymbol{H}$ and $\boldsymbol{Z}$, and then concatenates them together.

3.2.2. New Molecule Generation

After the learning phase, *TumFlow* can generate novel chemical structures conveying similar antitumoral activities to those in the training set and, using NN_{GI50}, it can also predict the antitumoral activity for each molecule. While this approach used alone proves highly valuable in other research domains, such as computer vision, where realistic face images need to be generated [70], it shows limitations when creating new antitumoral molecules. In fact, unlike scenarios where realistic faces are generated from datasets comprised of numerous real-world face images, the NCI-60 dataset includes many molecules with suboptimal antitumour efficacy. Consequently, employing this simplistic method to develop new molecules may yield structures with limited antitumour activities. *TumFlow* generates new molecules adopting a different approach that closely resembles structure optimization. Starting from a molecule with high antitumoral efficacy, through the use of the NN_{GI50} neural network, it modifies its structure to exhibit better antitumoral effects. Given a molecule in input, the optimization takes place many times, and for each of them *TumFlow* predicts the new molecule structure with associated predicted GI50 scores.

The optimization, which can also be seen in Figures 5 and S4, is performed using the gradient descent method, which is a pervasive optimization algorithm in machine learning employed to reduce the value of an objective function. Its primary objective is to iteratively approach the minimum of a function by moving in the direction of the most significant decrease in that function. In other words, *TumFlow* applies the gradient descent technique to the function NN_{GI50}, aiming to minimize the GI50 score. The optimization starts from a given molecule, i.e., the black ball in the top part of the image. Then, NN_{GI50} predicts the GI50 score associated with the starting molecules, as well as the direction to follow in the latent space to minimize the score. In other words, the direction to follow is the negative of the gradient returned by the NN_{GI50} component, as the lower the value is, the better the antitumoral properties are. Consequently, a new point in the latent space is selected, which can be decoded back to a molecule structure through $f_{\mathcal{E}}$ and $f_{\mathcal{V}|\mathcal{E}}$. This process is repeated several times. *TumFlow* performs the optimization process outlined above for each molecule, taking into account various gradient descent step values. In other words, for each optimization process, the movement performed in latent space involves making jumps of varying distance.

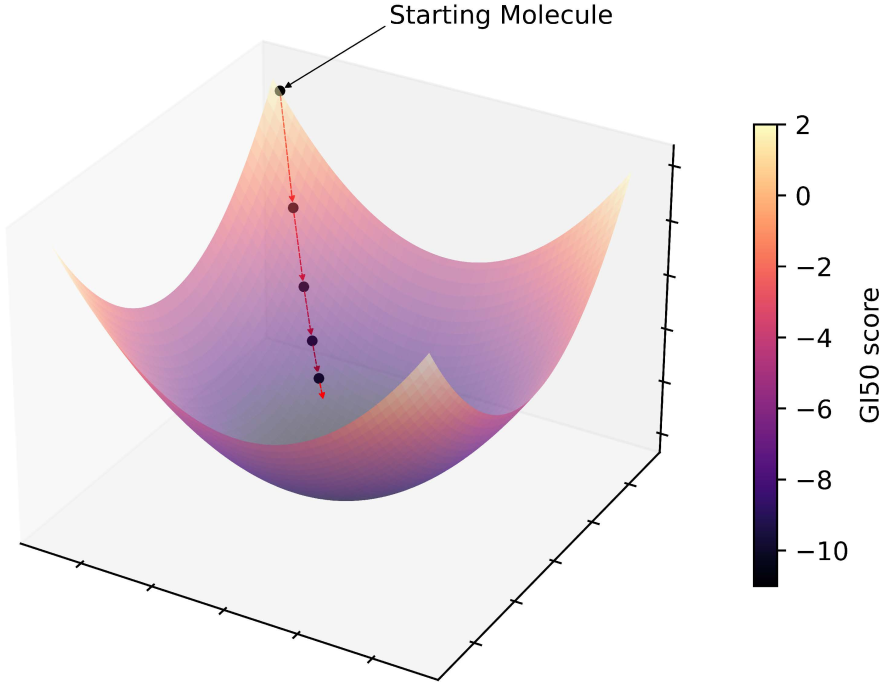

Figure 5. Figure representing the main idea behind the *TumFlow* generation of novel molecules with lower GI50 values, i.e., high antitumoral efficacy. Starting from an initial good molecule structure, new molecules are sampled in compliance with the gradient descent approach.

The generation of new molecules is performed following two different approaches:

(i) In the first approach, the starting point consists of molecules with higher antitumoral efficacy appearing in the training set, i.e., antitumoral molecules tested in vitro from the NCI-60 project;

(ii) In the second approach, the starting point consists of nine molecules, reported in Table S1, known for their efficacy in clinical treatments for melanoma.

In the process of generating new molecules, *TumFlow* is equipped with the ability to assess the Synthetic Accessibility Score (SAS) of drug-like molecules based on molecular complexity and fragment contributions [60]. This incorporation is essential because, even though *TumFlow* allows the generation of novel molecules, sometimes it produces energetically unstable structures and complex molecules that are challenging to synthesize. Further details are discussed in Section S7 of the Supplementary Material. In this work, the reported SAS values are normalized to fall within the range of $[0, 1]$, with lower values indicating greater ease of molecule synthesis and higher values suggesting increased difficulty in the synthesis process. The incorporation of this metric significantly enhanced the quality of molecules generated by *TumFlow*, facilitating the identification of compounds with potential antitumour effectiveness as well as a desired level of synthesis complexity.

4. Conclusions and Future Prospects

This investigation prominently showcases the generative capabilities and potential of *TumFlow* in oncological drug development. Unlike conventional methodologies, the presented approach harnesses the distinctive strengths of normalizing flow algorithms, notably their adeptness at modelling complex molecule distributions and generating accurate new data samples. This marks a substantial leap beyond traditional AI techniques, delivering unprecedented precision and efficiency in generating novel anticancer molecules.

In particular, this work demonstrates that *TumFlow* can identify crucial patterns and correlations between molecular structures and their predicted effectiveness against tumours like melanoma. It not only exhibits creativity in generating novel and promising molecules that have not been seen before but also has the capability to generate molecules not included in the training dataset, which already exist and have been subjected to in vivo experiments for antitumoral assessment. Although this creativity is essential for generating new drugs, there are some limitations that come with it, such as the feasibility of synthesis and the chemical instability of some generated structures.

By redefining the boundaries of possibilities within normalizing flow algorithms, the *TumFlow* model emerges not merely as a predictive instrument but as a designer, hopefully, of future oncological therapies. This methodology promises to diminish the time and financial constraints associated with drug discovery, steering researchers toward an era where swift, targeted, and potent cancer therapies are not merely conceivable but attainable. The *TumFlow* model, code, and documentation on GitHub [71] enhance reproducibility and accessibility in molecular generation. With scripts, trained weights, datasets, and a Docker container provided, it is a valuable resource for drug discovery research.

In summary, the application of the *TumFlow* model, as presented in this study, represents a significant advancement in the fight against cancer. This endeavour not only exemplifies the model's current achievements but also paves the way for a myriad of advancements in anticancer treatments and patient care.

Future research should consider various strategies to advance and increase the efficacy of the model presented in the context of oncological studies. One of these is the integration of broader and more diverse datasets, encompassing a wide variety of cancer types and molecules. Indeed, by exploiting drugs addressing different types of cancer, the model could learn complementary information that could enhance the discovery of new and more effective molecules. Moreover, given *TumFlow*'s tendency to generate energetically unstable complex structures, future works will consider the inclusion of the SAS values during the generation process and the inclusion of metrics to account for the energetic stability of the molecule. Additionally, incorporating information on the inhibition, lethality, and toxicity of antitumour molecules could provide an even more accurate algorithm for predicting new anticancer drugs. Another critical aspect is the continual enhancement of the computational and algorithmic capabilities of the model, to tackle challenges like interpreting molecular mechanisms and predicting drug side effects and resistance with the purpose of optimizing the molecules to yield better properties.

Supplementary Materials: References [72,73] are cited in the supplementary materials. The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms25116186/s1.

Author Contributions: Conceptualization, C.T.; formal analysis and methodology, D.R. and A.S.; software, D.R.; original draft preparation, D.R. and S.Y.; investigation, D.R., A.S. and C.T.; resources, D.R., S.M., A.S. and C.T.; data curation, D.R., S.M. and C.T.; writing—original draft preparation, D.R., S.Y. and C.T.; writing—review and editing, N.B., A.S. and S.M.; visualization, D.R., S.Y. and C.T.; supervision, N.B., A.S., S.M. and C.T.; funding acquisition, A.S., S.M. and C.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received financial support from BIRD213010/21, Department of Animal Medicine, Production and Health, University of Padua, and PRIN 2022, number 20227K7YJS CUP: C53D23004370006, Department of Pharmaceutical and Pharmacological Sciences, University of Padua. The authors acknowledge the support of the PNRR project FAIR—Future AI Research (PE00000013), under the NRRP MUR program funded by NextGenerationEU, the EuroHPC Joint Undertaking, for awarding access to Vega at IZUM, Slovenia, and the CINECA award under the ISCRA initiative for the availability of high-performance computing resources and support.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article and the Supplementary Materials. The code is openly available at https://github.com/drigoni/TumFlow (accessed on 6 February 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dzwierzynski, W.W. Melanoma risk factors and prevention. *Clin. Plast. Surg.* **2021**, *48*, 543–550. [CrossRef] [PubMed]
2. O'Neill, C.H.; Scoggins, C.R. Melanoma. *J. Surg. Oncol.* **2019**, *120*, 873–881. [CrossRef]
3. Gandini, S.; Sera, F.; Cattaruzza, M.S.; Pasquini, P.; Zanetti, R.; Masini, C.; Boyle, P.; Melchi, C.F. Meta-analysis of risk factors for cutaneous melanoma: III. Family history, actinic damage and phenotypic factors. *Eur. J. Cancer* **2005**, *41*, 2040–2059. [CrossRef]
4. Arnold, M.; de Vries, E.; Whiteman, D.C.; Jemal, A.; Bray, F.; Parkin, D.M.; Soerjomataram, I. Global burden of cutaneous melanoma attributable to ultraviolet radiation in 2012. *Int. J. Cancer* **2018**, *143*, 1305–1314. [CrossRef]
5. Erdei, E.; Torres, S.M. A new understanding in the epidemiology of melanoma. *Expert Rev. Anticancer. Ther.* **2010**, *10*, 1811–1823. [CrossRef] [PubMed]
6. Arioka, M.; Takahashi-Yanaga, F.; Kubo, M.; Igawa, K.; Tomooka, K.; Sasaguri, T. Anti-tumor effects of differentiation-inducing factor-1 in malignant melanoma: GSK-3-mediated inhibition of cell proliferation and GSK-3-independent suppression of cell migration and invasion. *Biochem. Pharmacol.* **2017**, *138*, 31–48. [CrossRef] [PubMed]
7. Chapman, P.B.; Hauschild, A.; Robert, C.; Haanen, J.B.; Ascierto, P.; Larkin, J.; Dummer, R.; Garbe, C.; Testori, A.; Maio, M.; et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. *N. Engl. J. Med.* **2011**, *364*, 2507–2516. [CrossRef]
8. Leach, D.R.; Krummel, M.F.; Allison, J.P. Enhancement of antitumor immunity by CTLA-4 blockade. *Science* **1996**, *271*, 1734–1736. [CrossRef]
9. Hodi, F.S.; O'day, S.J.; McDermott, D.F.; Weber, R.W.; Sosman, J.A.; Haanen, J.B.; Gonzalez, R.; Robert, C.; Schadendorf, D.; Hassel, J.C.; et al. Improved survival with ipilimumab in patients with metastatic melanoma. *N. Engl. J. Med.* **2010**, *363*, 711–723. [CrossRef]
10. Robert, C.; Long, G.V.; Brady, B.; Dutriaux, C.; Maio, M.; Mortier, L.; Hassel, J.C.; Rutkowski, P.; McNeil, C.; Kalinka-Warzocha, E.; et al. Nivolumab in previously untreated melanoma without BRAF mutation. *N. Engl. J. Med.* **2015**, *372*, 320–330. [CrossRef]
11. Zang, C.; Wang, F. Moflow: An invertible flow model for generating molecular graphs. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Virtual, 23–27 August 2020; pp. 617–626.
12. Harrer, S.; Shah, P.; Antony, B.; Hu, J. Artificial intelligence for clinical trial design. *Trends Pharmacol. Sci.* **2019**, *40*, 577–591. [CrossRef] [PubMed]
13. Mullard, A. 2021 FDA approvals. *Nat. Rev. Drug Discov.* **2022**, *21*, 83–88. [CrossRef] [PubMed]
14. Statista. Spending of the U.S. Pharmaceutical Industry on Research and Development at Home and Abroad from 1990 to 2022 (in Million U.S. Dollars). *In Statista*. 2023. Retrieved 1 January 2024. Available online: https://www.statista.com/statistics/265090/us-pharmaceutical-industry-spending-on-research-and-development/ (accessed on 6 February 2024).
15. Jiang, F.; Jiang, Y.; Zhi, H.; Dong, Y.; Li, H.; Ma, S.; Wang, Y.; Dong, Q.; Shen, H.; Wang, Y. Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* **2017**, *2*, 230–243. [CrossRef] [PubMed]
16. Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **2019**, *18*, 463–477. [CrossRef] [PubMed]

17. Hassanzadeh, P.; Atyabi, F.; Dinarvand, R. The significance of artificial intelligence in drug delivery system design. *Adv. Drug Deliv. Rev.* **2019**, *151*, 169–190. [CrossRef] [PubMed]
18. Fakoor, R.; Ladhak, F.; Nazi, A.; Huber, M. Using deep learning to enhance cancer diagnosis and classification. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 6–21 June 2013; ACM: New York, NY, USA, 2013; Volume 28, pp. 3937–3949.
19. Munir, K.; Elahi, H.; Ayub, A.; Frezza, F.; Rizzi, A. Cancer diagnosis using deep learning: A bibliographic review. *Cancers* **2019**, *11*, 1235. [CrossRef] [PubMed]
20. Bloice, M.D.; Roth, P.M.; Holzinger, A. Biomedical image augmentation using Augmentor. *Bioinformatics* **2019**, *35*, 4522–4524. [CrossRef]
21. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014.
22. Kusner, M.J.; Paige, B.; Hernández-Lobato, J.M. Grammar variational autoencoder. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 1945–1954.
23. Dai, H.; Tian, Y.; Dai, B.; Skiena, S.; Song, L. Syntax-Directed Variational Autoencoder for Structured Data. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.
24. Liu, Q.; Allamanis, M.; Brockschmidt, M.; Gaunt, A. Constrained graph variational autoencoders for molecule design. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 7806–7815.
25. Ma, T.; Chen, J.; Xiao, C. Constrained generation of semantically valid graphs via regularizing variational autoencoders. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 7113–7124.
26. Jin, W.; Barzilay, R.; Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2323–2332.
27. Rigoni, D.; Navarin, N.; Sperduti, A. Conditional constrained graph variational autoencoders for molecule design. In Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra Australia, 1–4 December 2020; IEEE: New York, NY, USA, 2020; pp. 729–736.
28. Rigoni, D.; Nicolo, N.; Alessandro, S. A Systematic Assessment of Deep Learning Models for Molecule Generation. In Proceedings of the ESANN 2020-Proceedings, 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 2–4 October 2020; pp. 547–552.
29. Rigoni, D.; Navarin, N.; Sperduti, A. RGCVAE: Relational Graph Conditioned Variational Autoencoder for Molecule Design. *arXiv* **2023**, arXiv:2305.11699.
30. Hy, T.S.; Kondor, R. Multiresolution equivariant graph variational autoencoder. *Mach. Learn. Sci. Technol.* **2023**, *4*, 015031. [CrossRef]
31. Bhadwal, A.S.; Kumar, K.; Kumar, N. NRC-VABS: Normalized Reparameterized Conditional Variational Autoencoder with applied beam search in latent space for drug molecule design. *Expert Syst. Appl.* **2024**, *240*, 122396. [CrossRef]
32. De Cao, N.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *arXiv* **2018**, arXiv:1805.11973.
33. Tsujimoto, Y.; Hiwa, S.; Nakamura, Y.; Oe, Y.; Hiroyasu, T. L-MolGAN: An improved implicit generative model for large molecular graphs. *ChemRxiv* **2021**, chemrxiv.14569545.
34. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
35. Shi, C.; Xu, M.; Zhu, Z.; Zhang, W.; Zhang, M.; Tang, J. Graphaf: A flow-based autoregressive model for molecular graph generation. *arXiv* **2020**, arXiv:2001.09382.
36. Madhawa, K.; Ishiguro, K.; Nakago, K.; Abe, M. GraphNVP: An Invertible Flow-Based Model for Generating Molecular Graphs. *arXiv* **2019**, arXiv:1905.11600.
37. Kobyzev, I.; Prince, S.J.; Brubaker, M.A. Normalizing flows: An introduction and review of current methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3964–3979. [CrossRef]
38. Kuznetsov, M.; Polykovskiy, D. MolGrow: A graph normalizing flow for hierarchical molecular generation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 8226–8234.
39. Faez, F.; Ommi, Y.; Baghshah, M.S.; Rabiee, H.R. Deep graph generators: A survey. *IEEE Access* **2021**, *9*, 106675–106702. [CrossRef]
40. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 6 July–11 July 2015; pp. 2256–2265.
41. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
42. Austin, J.; Johnson, D.D.; Ho, J.; Tarlow, D.; Van Den Berg, R. Structured denoising diffusion models in discrete state-spaces. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17981–17993.
43. Nichol, A.Q.; Dhariwal, P. Improved denoising diffusion probabilistic models. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 8162–8171.
44. Vignac, C.; Krawczuk, I.; Siraudin, A.; Wang, B.; Cevher, V.; Frossard, P. Digress: Discrete denoising diffusion for graph generation. *arXiv* **2022**, arXiv:2209.14734.
45. Luo, T.; Mo, Z.; Pan, S.J. Fast graph generation via spectral diffusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *46*, 3496–3508. [CrossRef] [PubMed]

46. Jo, J.; Lee, S.; Hwang, S.J. Score-based generative modeling of graphs via the system of stochastic differential equations. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MD, USA, 17–23 July 2022; pp. 10362–10383.
47. Huang, H.; Sun, L.; Du, B.; Fu, Y.; Lv, W. Graphgdp: Generative diffusion processes for permutation invariant graph generation. In Proceedings of the 2022 IEEE International Conference on Data Mining (ICDM), Orlando, FL, USA, 28 November–1 December 2022; IEEE: New York, NY, USA, 2022; pp. 201–210.
48. Xu, M.; Yu, L.; Song, Y.; Shi, C.; Ermon, S.; Tang, J. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv* **2022**, arXiv:2203.02923.
49. Hoogeboom, E.; Gritsenko, A.A.; Bastings, J.; Poole, B.; Berg, R.v.d.; Salimans, T. Autoregressive diffusion models. *arXiv* **2021**, arXiv:2110.02037.
50. Huang, H.; Sun, L.; Du, B.; Lv, W. Conditional diffusion based on discrete graph structures for molecular graph generation. *arXiv* **2023**, arXiv:2301.00427. [CrossRef]
51. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
52. Mazuz, E.; Shtar, G.; Shapira, B.; Rokach, L. Molecule generation using transformers and policy gradient reinforcement learning. *Sci. Rep.* **2023**, *13*, 8799. [CrossRef]
53. Bagal, V.; Aggarwal, R.; Vinod, P.; Priyakumar, U.D. MolGPT: Molecular generation using a transformer-decoder model. *J. Chem. Inf. Model.* **2021**, *62*, 2064–2076. [CrossRef]
54. Rothchild, D.; Tamkin, A.; Yu, J.; Misra, U.; Gonzalez, J. C5t5: Controllable generation of organic molecules with transformers. *arXiv* **2021**, arXiv:2108.10307.
55. Dollar, O.; Joshi, N.; Beck, D.A.; Pfaendtner, J. Attention-based generative models for de novo molecular design. *Chem. Sci.* **2021**, *12*, 8362–8372. [CrossRef]
56. Huang, C.W.; Krueger, D.; Lacoste, A.; Courville, A. Neural autoregressive flows. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2078–2087.
57. Ohno, H. Training data augmentation: An empirical study using generative adversarial net-based approach with normalizing flow models for materials informatics. *Appl. Soft Comput.* **2020**, *86*, 105932. [CrossRef]
58. Dinh, L.; Sohl-Dickstein, J.; Bengio, S. Density estimation using Real NVP. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
59. Kingma, D.P.; Dhariwal, P. Glow: Generative flow with invertible 1×1 convolutions. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 10236–10245.
60. Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminform.* **2009**, *1*, 1–11. [CrossRef] [PubMed]
61. Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754. [CrossRef] [PubMed]
62. NCI-60 Project. Available online: https://dtp.cancer.gov/discovery_development/nci-60/ (accessed on 1 October 2023).
63. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. [CrossRef]
64. Weininger, D.; Weininger, A.; Weininger, J.L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101. [CrossRef]
65. Weininger, D. SMILES. 3. DEPICT. Graphical depiction of chemical structures. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 237–243. [CrossRef]
66. PubChem. Available online: https://pubchem.ncbi.nlm.nih.gov/ (accessed on 10 April 2024).
67. Docker. Available online: https://www.docker.com/ (accessed on 22 May 2024).
68. National Cancer Institute (NCI). Available online: https://dtp.cancer.gov/ (accessed on 1 October 2023).
69. Schlichtkrull, M.; Kipf, T.N.; Bloem, P.; Van Den Berg, R.; Titov, I.; Welling, M. Modeling relational data with graph convolutional networks. In Proceedings of the The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, 3–7 June 2018; Proceedings 15. Springer: Berlin/Heidelberg, Germany, 2018; pp. 593–607.
70. Wu, X.; Zhang, Q.; Wu, Y.; Wang, H.; Li, S.; Sun, L.; Li, X. F^{3}A-GAN: Facial Flow for Face Animation With Generative Adversarial Networks. *IEEE Trans. Image Process.* **2021**, *30*, 8658–8670. [CrossRef]
71. GitHub. Available online: https://github.com/ (accessed on 22 May 2024).
72. RDKit. Available online: https://www.rdkit.org/ (accessed on 22 May 2024).
73. Dinh, L.; Krueger, D.; Bengio, Y. NICE: Non-linear Independent Components Estimation. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.