Stereochemistry-aware string-based molecular generation

Gary Tom,[1,2] Edwin Yu,[1] Naruki Yoshikawa,[2,3] Kjell Jorner,[1,3,4,5,*] and Alán Aspuru-Guzik[1,3,2,6,7,8,†]

[1] *Department of Chemistry, University of Toronto, Canada.*
[2] *Vector Institute for Artificial Intelligence, Toronto, Canada.*
[3] *Department of Computer Science, University of Toronto, Canada.*
[4] *Department of Chemistry and Chemical Engineering, Chalmers University of Technology, Sweden.*
[5] *Institute of Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zurich, Switzerland*
[6] *Department of Chemical Engineering & Applied Chemistry, University of Toronto, Canada.*
[7] *Department of Materials Science & Engineering, University of Toronto, Canada.*
[8] *Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR), Canada.*

This study investigates the impact of incorporating stereochemical information, a crucial aspect of computational drug discovery and materials design, in molecular generative modelling. We present a comprehensive comparison of stereochemistry-aware and conventionally stereochemistry-unaware string-based generative approaches, utilizing both genetic algorithms and reinforcement learning-based techniques. To evaluate these models, we introduce novel benchmarks specifically designed to assess the importance of stereochemistry-aware generative modelling. Our results demonstrate that stereochemistry-aware models generally perform on par with or surpass conventional algorithms across various stereochemistry-sensitive tasks. However, we also observe that in scenarios where stereochemistry plays a less critical role, stereochemistry-aware models may face challenges due to the increased complexity of the chemical space they must navigate. This work provides insights into the trade-offs involved in incorporating stereochemical information in molecular generative models and offers guidance for selecting appropriate approaches based on specific application requirements.

I. Introduction

Generative models have become increasingly prominent in the fields of inverse design and molecular discovery, offering a computational approach to explore vast chemical spaces efficiently [1–10]. These models employ machine learning techniques to generate novel molecular structures with targeted properties, potentially expediting the traditionally lengthy and resource-intensive process of molecular design [11–14]. Generative models can propose new and potentially viable compounds, adhering to specified criteria. Methods such as genetic algorithms define heuristics for exploring the space of chemicals, while deep-learning methods learn the chemical space distribution from molecular databases. The literature presents a diverse array of approaches in this domain, including but not limited to variational autoencoders (VAEs) [1, 15–18], generative adversarial networks (GANs) [4, 19, 20], reinforcement learning (RL) [21–26], genetic algorithms (GAs) [27–30], and transformer-based architectures [31–34]. These methodologies have demonstrated utility across various applications in drug discovery and materials science, facilitating rapid *in silico* screening and optimization of molecular structures [35].

The evaluation and benchmarking of generative models for molecular discovery initially focused on determining the goodness of the reproduction of the structures in the dataset chemical space—generation not conditioned on the functional properties of the molecules. These metrics typically emphasize distribution learning, examining the model's ability to capture and reproduce the underlying distribution of the training data [3, 36–38]. Other evaluation criteria include the (1) novelty of generated molecules, which measures the proportion of unique structures not present in the training set; (2) diversity, which assesses the structural variation among the generated molecules; and validity, which ensures that the proposed structures adhere to chemical feasibility constraints (e.g., valid Lewis structures and valency constraints) [18, 39].

While these metrics provide insights into a model's generative capabilities, there is a growing recognition of the need for more realistic and task-specific benchmarks [40–43]. The emphasis on general distribution learning, while important, may not fully capture the model's performance in addressing specific chemical challenges. Additionally, performances on task-oriented benchmarks based on simple heuristic fitness functions, such as penalized log water-octanol partition coefficient [1, 44], similarity/rediscovery tasks [18, 39], or quantitative estimate of drug-likeliness (QED) [45], are handily maximized by modern generative models [46–48], and even trivially satisfied by randomly inserting carbon atoms into the molecules [49]. These simplistic fitness functions often fail to capture chemical constraints, allowing models to exploit failure modes by reward hacking, and generate molecules with high scores but undesirable properties, such as chemical instability or synthetic infeasibility [50, 51]. As the field advances, there is an increasing demand for benchmarks that are more closely aligned with real-world applications in drug discovery, materials design, and other domains of chemistry [52]. This

* kjell.jorner@chem.ethz.ch
† alan@aspuru.com

shift towards more targeted evaluation methods would provide a more nuanced and practically relevant assessment of generative models, potentially accelerating their adoption and impact in real-world molecular discovery scenarios.

Despite the advances in generative models for molecular design, the incorporation of stereochemical information remains a significant challenge. Molecular stereochemistry, the 3D arrangement of atoms within a molecule, significantly influences its chemical properties and biological activity [53]. Many current methods either ignore stereochemistry entirely or consider it as a post-processing step after molecule generation. This approach is suboptimal, as stereochemistry plays a crucial role in determining a molecule's properties and biological activity. The importance of stereochemistry is particularly evident in drug discovery, where the spatial arrangement of atoms can significantly influence a compound's pharmacological properties [54, 55]. Properties such as binding affinity to target proteins, metabolic stability, and toxicity can be profoundly affected by stereochemistry. For example, the synthesis of methadone produces racemic mixtures of enantiomers—molecules that are mirror images of each other—*R*-methadone, and *S*-methadone. While *R*-methadone acts as an opioid for pain relief, *S*-methadone has been identified to bind to the hERG protein and can lead to severe side-effects, such as heart attacks or cardiac arrest [56]. In materials science, stereochemistry can impact crystal packing, optical properties, synthesis, and reactivity [57–60]. By not explicitly accounting for stereochemistry during the generative process, models may overlook critical aspects of molecular behaviour, potentially leading to inefficiencies in the discovery pipeline and missed opportunities for identifying optimal candidates for a given application.

In our work, we study the effects of stereochemistry on string-based generative models. We evaluate the models, both with and without stereochemistry-awareness, on a variety of molecular design tasks that are sensitive to the stereochemistry of molecules. Additionally, we explore different string representations of molecular graphs, and create a workflow for benchmarking the models, which includes a novel fitness function based on the circular dichroism spectra of molecules. We find that stereo-aware models perform as well as, or better than non-stereo models, but the performance increase of the stereo models are dependent on the sensitivity of the task to stereochemistry. The models and the fitness functions are all available at https://github.com/aspuru-guzik-group/stereogeneration.

II. Methods

To study the effects of stereochemistry on molecular generative models, we implement RL and GA methods, which have been shown to be strong baselines for molecular generation tasks [24, 30, 40]. We modify the `REINVENT` [21, 22] and `JANUS` [29] models to permit the representation of stereochemical information. In these models, the molecular graphs are represented as strings, where `REINVENT` uses Simplified Molecular-Input Line-Entry System (SMILES) [61], and `JANUS` uses SELF-Referencing Embedded Strings (SELFIES) [62], or GroupSELFIES [63]. We choose to use string-based generative models due to their expressiveness and flexibility in exploring chemical space when compared to graph-based methods [37], and their native support of stereochemical string tokens. By directly comparing the models with and without the stereochemistry-awareness across the various tasks, we can elucidate the effect of stereochemistry in the molecular generation process.

A. Stereochemistry

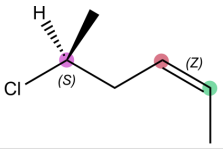
We focus on two primary forms of stereoisomerism: E/Z geometric diastereomers, arising from restricted rotation around double bonds, and R/S diastereomers and enantiomers, determined by the arrangement of substituents around chiral centres. Enantiomers are non-superimposable mirror images of each other and often have different optical activity and physical properties. Diastereomers, stereoisomers that are not mirror images of each other, also often exhibit different physical and chemical properties.

While we incorporate E/Z and R/S isomerism, we do not explicitly account for axial chirality, a type of chirality arising from hindered rotation around single bonds [64], or ring isomers. This omission limits our model's ability to generate and differentiate atropisomers, a specific class of axially chiral molecules.

B. String representations

SMILES were initially created as a compact representation of molecular graphs for purposes of database retrieval, and substructure searching. When used in generative models, SMILES of generated molecules can sometimes violate the grammar of the representation, resulting in invalid SMILES. To address this, SELFIES made use of overloaded tokens, and local definitions of rings and branches to create a robust representation that will always translate to a valid molecular graph. GroupSELFIES further extended SELFIES by allowing for custom tokens which can encode groups with specified attachment points. For more details on string representations of molecules, we direct the readers to Krenn *et al.* [65]. We also note that there are other string representations that incorporate stereochemistry which are not explored in this work [66].

All three representations natively encode stereochemi-



SMILES:	C/C=C\C[C@H](C)Cl
SELFIES:	[C][/C][=C][\C][C@H1][Branch1][C][C][Cl]
Group SELFIES:	[C][/C][=C][\C][:0chiral][Ring1][C][pop] [Ring1][Cl][pop]

FIG. 1. **Example of isomeric molecule encoded with SMILES, SELFIES, and GroupSELFIES**. Stereoinformation is labelled and highlighted in the structure and in the corresponding stereochemical tokens in each representation. Note that SELFIES tokens are denoted by `[·]`.

cal information (Figure 1). SMILES encode counterclockwise and clockwise chirality with "`@`" and "`@@`" tokens, respectively. *E-Z* stereoisomers are denoted with "\" and "/" before the characters to indicate the position of a bond relative to an adjacent double bond. The same characters are used in the SELFIES stereochemical tokens, while also maintaining the robustness of the representation. GroupSELFIES defines *E-Z* stereoisomers in the same way as SMILES and SELFIES, but defines chirality through unique tokens for each chiral centre and for all possible attachment points. The attachment points directly encode the chirality of the chiral centre, with different attachment indices in the tokens specifying the order of substituents around the chiral centre.

For all experiments, we use a subset of the ZINC15 database that was randomly sub-sampled by Gomez-Bombarelli et al. [1, 67]. This dataset is composed of about 250,000 commercially available drug-like molecules. Stereoinformation is defined for most molecules in the dataset. Any molecules with ambiguous stereochemistry are assigned stereochemistry by randomly selecting from a list enumerating all unspecified stereo-centres using RDKit cheminformatics software [68]. For the non-stereo experiments, the stereoinformation is discarded, and duplicates resulting from the loss of stereoinformation are removed. Subsequently, the unique string tokens are collected to create an alphabet, with stereo and non-stereo alphabets for each representation. The GroupSELFIES representation has an additional essential set of chiral tokens, which are appended to the alphabet generated from the dataset.

C. Generative models

`REINVENT` is an RL algorithm that uses a recurrent neural network (RNN) pretrained on a dataset of SMILES as a chemical language model agent [21, 22, 24]. When provided a token from a SMILES string, the RNN is trained to generate a conditional distribution of the subsequent tokens in the sequence. A memory state is passed into the model as well, retaining information about previous tokens of the sequence observed by the model. The RNN is first pretrained on the initial ZINC dataset, allowing it to learn the grammar of the stereo and non-stereo SMILES in the dataset, producing 94% and 91% average validity of generated SMILES, respectively. During the RL optimization, the prior RNN is fine-tuned after each generation by a loss function augmented by the fitness score achieved by the molecule $S \in [0, 1]$, with good candidates scoring $S = 1$, and poor candidates and invalid SMILES scoring $S = 0$. With each iteration, the RL algorithm will aim to optimize the molecules to maximize the fitness function. Note that SELFIES can also be used with `REINVENT`, but previous studies have demonstrated that the RNN model is sufficiently capable of generating valid SMILES, and no significant performance gain is observed for SELFIES-`REINVENT` [40].

On the other hand, `JANUS` admits only SELFIES-based representations. Leveraging the robustness of SELFIES representation, `JANUS` can perform mutation and crossover operations, as defined in the STONED algorithm [69]. `JANUS` maintains two separate populations for exploration and exploitation of chemical space. The exploration set is generated by mutation and crossover operations within the entire population, while the exploitation set is generated through a series of mutations on the fittest molecules. The best candidates found in the exploitation set are then exchanged with the worst candidates in the exploration set. At each iteration, selection pressure from the fitness function allows the model to converge toward the optimum.

In our workflow, we implement the GroupSELFIES version of `JANUS`, dubbed `GroupJANUS`, which operates in the same fashion as `JANUS`. In order to isolate the effect of the stereochemical tokens, only the chiral group tokens are used in `GroupJANUS`; no other groups are encoded in the GroupSELFIES grammar. For both `JANUS` and `GroupJANUS`, the mutation operations depend on the random sampling of tokens in the alphabet. For both models, the inclusion of stereochemical tokens greatly increases the size of the alphabet, and structural tokens which are responsible for encoding molecular rings and branches are less likely to be sampled. To account for this imbalance, structural tokens—such as `[RingX]`, `[BranchX]` and the GroupSELFIES specific `[pop]` tokens—are weighted such that they are sampled with the same probability as in the non-stereo alphabet.

D. Experiments

We perform three stereochemistry-sensitive generative experiments to benchmark the models. We study `REINVENT`, `JANUS`, and `GroupJANUS` with SMILES, SELFIES, and GroupSELFIES representations, respectively. While the stereo models will output specific stereoisomers, non-stereo models are unable to distinguish be-

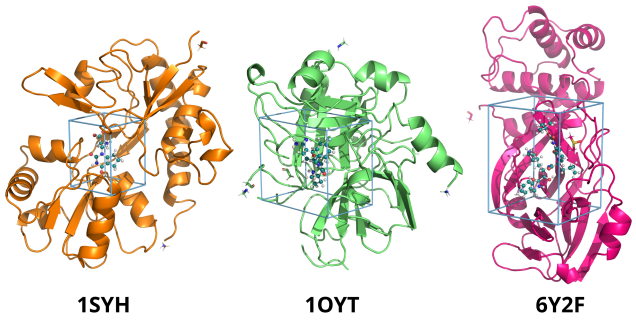


FIG. 2. **Structures of proteins with native ligands.** The structures are from the Protein Data Bank [71]. The native ligand in the binding pocket is shown inside a bounding box.

tween different isomers with the same molecular graph connectivity. Non-stereo SMILES are randomly assigned stereochemical information before fitness evaluation, ensuring that stereochemical information is assigned but not passed to the generative model.

1. *Stereoisomer rediscovery task*

Rediscovery tasks in molecular generative modelling benchmarking aim to evaluate a model's ability to recreate the structure of known molecules. The structural similarity is measured by the Tanimoto similarity of molecular fingerprints—typically, extended circular fingerprints (ECFPs), bit vectors based on the topological features of a certain radius in the molecular graph [70]. The model successfully rediscovers a target when the similarity is 1.0. While rediscovery tasks are not useful in practice, since the target molecules are known *a priori*, they serve as useful baselines to study the generative capabilities of the models in directly optimizing molecular structures, rather than chemical function. Previous benchmarking rediscovery tasks ignore the stereochemistry of the molecular structures [39]. We include the stereoinformation as part of the target through the use of isomeric ECFPs. For this, we chose to perform rediscovery of (R)-albuterol (used in asthma treatment) and mestranol (used as estrogen medication for hormone therapy), with one and five chiral centres, respectively.

2. *Protein-ligand docking task*

Protein-ligand interactions are associated with the bioactivity of drug molecules. Ligands are molecules that bind inside the protein binding pockets, forming intermolecular interactions with the amino acids of the protein, activating or inhibiting biological functions of the protein. For the benchmark, we use the high-throughput docking score implemented in the `Tartarus` benchmark [43], which uses the `smina` software to simulate the protein-ligand binding affinity [72].

Because the scoring function takes in a 3D conformer of the molecules, a conformer search is performed using RDKit to find the lowest energy conformer, respecting all specified stereoinformation, followed by energy relaxation with the Merck Molecule Force Field 94 (MMFF94) [73]. The molecule is placed inside the binding pocket to sample binding poses; the resulting docking score is maximized. The binding pocket is defined as the bounding box encompassing the volume occupied by the protein's native ligand with 3Å padding.

We perform the protein-ligand docking task for three different targets, visualized in Figure 2 with their respective bounding boxes. Both 1SYH and 6Y2F are targets from `Tartarus`: 1SYH is associated with neurological diseases, and 6Y2F is responsible for the translation of the SARS-CoV-2 virus RNA. We also include the 1OYT protein, which is associated with blood coagulation [74], and has a binding pocket with a volume between those of 1SYH and 6Y2F.

3. *Circular dichroism task*

We finally developed a task based on circular dichroism (CD), which directly probes the chirality of structures, making it the ideal task for studying the effects of stereoinformation in molecular generation. CD produces spectra of the absorption of left- and right-handed polarized light in chemical species, and can be used to study folding structures in proteins [75], or chiral optical properties of materials, which have light manipulation and photonics applications [57, 76–78].

In this task, like before, the molecules are 3D embedded with RDKit. The conformer search and geometry optimization is performed using `crest` [79–81] and semi-empirical extended tight-binding (xTB) [82] at the GFN2 level of theory [83].

The xTB calculation quickly produces orbitals and orbital energies, which can be treated using simplified Tamm-Dancoff approximated (sTDA) time-dependent density functional theory (TD-DFT). This workflow, `sTDA-xTB`, produces CD spectra of the lowest energy conformers relatively quickly, even for molecular systems with hundreds of atoms [84, 85]. A peak score is defined as the signed area under the spectrum for wavelengths 450–550nm, a region where small organic molecules can have CD signals, and is also within the visible range for possible materials applications. Maximizing the peak score produces chiral optically active materials within the blue region of visible light.

III. Results

We evaluate optimization performance by looking at the optimization trace, which plots the cumulative top-1

score achieved as a function of the generation of the campaign. We do this across the models, stereo and non-stereo aware, on the aforementioned tasks. For all tasks, the models are allotted 1000 fitness oracle calls. The GAs run 50 generations with population size of 200 molecules, while REINVENT runs 100 generations of 100 molecules to allow for more policy updates throughout the optimization. Experiments were repeated with 10 times, and statistical significance was determined by Student t-test.

Additionally, we use the area-under-curve (AUC) of the optimization traces as a quantitative measure of the optimization performance. For the AUC calculation, the number of generations is normalized from 0 to 1. For the rediscovery tasks, the similarity score and the AUC are both bounded by 0 and 1. For the docking and CD tasks, there is no maximum achievable score. Therefore, we normalize the AUC scores by the best score in the initial ZINC dataset. Higher AUC indicates the generation of higher scoring molecules, and also earlier discovery of such molecules. The AUC scores are found in Table I.

	Tasks	REINVENT	JANUS	GroupJANUS
Non-stereo	(R)-albuterol rediscovery	0.487 ± 0.058	0.790 ± 0.105	0.840 ± 0.109
	Mestranol rediscovery	0.292 ± 0.034	0.633 ± 0.031	0.672 ± 0.032
	1SYH docking	0.900 ± 0.020	1.033 ± 0.031	1.084 ± 0.053
	1OYT docking	0.954 ± 0.013	1.064 ± 0.028	1.068 ± 0.028
	6Y2F docking	0.987 ± 0.015	1.068 ± 0.052	1.067 ± 0.029
	CD spectral peak score	0.413 ± 0.117	2.007 ± 0.352	2.066 ± 0.761
Stereo	(R)-albuterol rediscovery	0.403 ± 0.053	**0.931 ± 0.044**	**0.923 ± 0.035**
	Mestranol rediscovery	0.280 ± 0.032	**0.843 ± 0.087**	**0.918 ± 0.074**
	1SYH docking	0.887 ± 0.011	**1.065 ± 0.031**	1.106 ± 0.070
	1OYT docking	0.940 ± 0.021	**1.099 ± 0.027**	1.059 ± 0.035
	6Y2F docking	0.979 ± 0.023	1.088 ± 0.043	1.065 ± 0.042
	CD task	0.385 ± 0.111	**2.884 ± 1.009**	2.198 ± 0.563

TABLE I. **AUC of optimization traces for all tasks, for stereo and non-stereo aware models.** The mean and standard deviation are reported. Statistically significantly better (higher) AUC scores between the non-stereo and stereo variants are bolded.

A. Stereoisomer rediscovery task

The optimization traces for the rediscovery tasks are shown in Figure 3. The higher number of chiral centres in mestranol make it a more difficult target for rediscovery. This is clearly shown in the optimization traces of the REINVENT models—mestranol rediscovery does not achieve similarity higher than the initial dataset. When compared to rediscovery in other studies [26, 39, 40], REINVENT optimization performance is greatly reduced when stereochemistry is introduced. There are no statistically significant differences in the performance of REINVENT when comparing stereo and non-stereo models.

The stereo-aware JANUS and GroupJANUS models significantly outperform the non-stereo-aware models, indicating the ability of the stereo model in learning specific stereochemistries in molecular structures. For (R)-albuterol, both stereo GAs successfully rediscover the

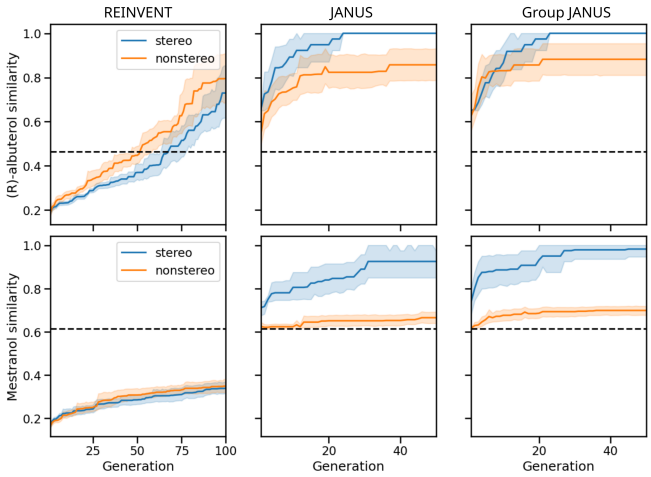


FIG. 3. **Optimization traces for rediscovery tasks.** The cumulative top-1 similarity score to the target molecule as a function of generation of optimization. Shaded regions indicate the 95% confidence interval. The dashed line is the best score found in the starting dataset.

structure for all runs, with no significant differences between JANUS and GroupJANUS. For mestranol, the use of GroupSELFIES slightly improves the optimization for the stereo-aware model, when compared to JANUS with SELFIES.

B. Protein-ligand docking task

Moving beyond simple structural reproduction, the protein-ligand docking task assesses the practical utility of generative models in a drug discovery context. The optimization traces for the docking tasks are in figure 4. We again observe that REINVENT struggles to improve upon the results of the ZINC dataset, with the exception of the 6Y2F protein. There are no differences between the stereo and non-stereo variants of REINVENT.

Meanwhile, both GAs optimize better than REINVENT. For JANUS, we observe consistent improvements in optimization performance with stereo-aware models for generating ligands for 1SYH and 1OYT. The faster optimization of the stereo GAs are also reflected in the AUC score (Table I). In the case of the 6Y2F target, possessing a comparatively larger and more flexible binding pocket, the difference in performance between stereo and non-stereo models was less pronounced. This observation implies that for certain targets, the impact of stereochemistry on binding affinity might be less critical, with other molecular features playing a dominant role.

C. Circular dichroism task

The results of CD peak score optimization task are shown in Figure 5. There are no differences between the non-

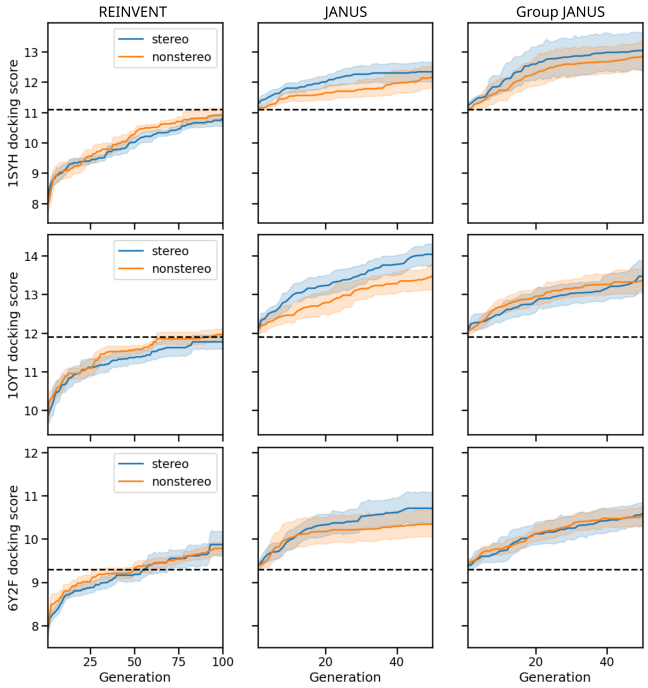


FIG. 4. **Optimization traces for docking tasks.** The cumulative top-1 docking score for protein targets as a function of generation of optimization. Shaded regions indicate the 95% confidence interval. The dashed line is the best score found in the starting dataset.

stereo and stereo `REINVENT`, which are unable to improve upon the scores of the initial dataset. The `GroupJANUS` optimization also shows no difference between the stereo and non-stereo models. However, the stereo-aware `JANUS` is capable of generating molecules with stronger CD signals than the non-stereo-aware counterpart. The results indicate that the CD task is a suitably stereochemistry-sensitive optimization task for molecular generative modelling.

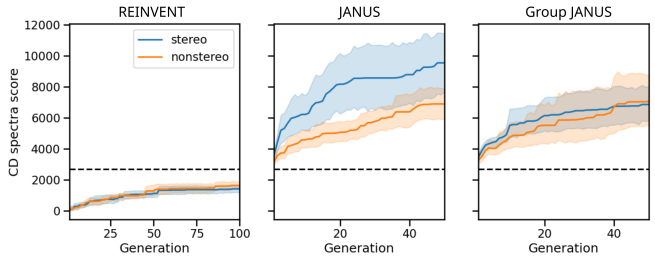


FIG. 5. **Optimization traces for CD task.** The cumulative top-1 CD peak score as a function of generation of optimization. Shaded regions indicate the 95% confidence interval. The dashed line is the best score found in the starting dataset.

IV. Discussion

While it may seem intuitive that stereochemistry-aware models would be capable of generating better molecules for optimization tasks, the optimization performance of generative models depends on the stereochemistry-sensitivity of the task. For explicit optimization of molecular structures, stereo-aware GAs perform better than non-stereo counterparts. In the docking task, we observe that stereo GAs boost the optimization performance for 1SYH and 1OYT. In the case of 6Y2F, the generated ligand molecules are larger, in order to fit in the bigger protein binding pocket. Larger structural changes such as additional of fragments and functional groups allow the models to more quickly traverse the permitted molecule space, while slight changes in stereochemistry only result in small changes in the docking score. In these tasks, stereo models still perform as well as non-stereo models. The CD spectra task directly probes the effects of chirality, and the spectra is less related to specific molecular size or functional groups. In this task, stereo `JANUS` outperforms non-stereo `JANUS`.

Except for the rediscovery tasks, unlike `JANUS`, the stereo and non-stereo variants of `GroupJANUS` perform similarly. This may be due to inefficiencies of the GroupSELFIES representation of stereochemistry. The addition of stereochemistry tokens increase the alphabet size by almost 3 times. Also, all additional tokens and group tokens are overloaded to ensure robustness. We hypothesize that the increased number of tokens interferes with the decoding of stereoisomeric GroupSELFIES, truncating molecules at rings and branches.

Additionally, GAs perform better than the `REINVENT` model for the same number of oracle calls, results which are consistent with previous studies [30]. The evolutionary approach of GAs will always select the members of the population that maximize the fitness, meaning the GA cannot perform worse than the previous (or initial) generations. GAs are also not encumbered by the prior chemical space distribution of the training set, unlike deep learning methods like `REINVENT`, allowing the generation of more diverse molecules. Due to the prior model, the `REINVENT` agent requires more oracle calls and more frequent retraining to condition for higher rewards. In this study, we limit the models to 1000 oracle calls, as opposed to 5000 in [43] or 10,000 in [40]. No improvement was observed for stereo `REINVENT`. The expanded alphabet increases the number of possible actions the RL, requiring the model to learn a more complex policy which may have impeded stereo `REINVENT`.

V. Conclusion

This study presents a comprehensive investigation into the incorporation of stereochemical information within molecular generative models, focusing on established

techniques such as string-based RL and GAs. We aim to provide a nuanced understanding of the impact of stereochemistry awareness by employing a suite of evaluation metrics, including both conventional benchmarks and newly designed tasks specifically tailored to assess the role of stereochemistry. A key contribution of this work is the introduction of a novel CD-based task, which proved to be suitable for probing the effects of chirality in the generated molecules.

Our findings highlight the importance of considering task-specific requirements when deciding whether to include stereochemical information within the generative process. In cases where different stereoisomers can significantly influences the desired molecular properties, the inclusion of stereochemistry led to improved performance. Specifically, we observed that stereochemistry-aware GA `JANUS` consistently outperformed their non-stereo counterparts in generating molecules for stereoisomer rediscovery, docking to proteins 1SYH and 1OYT, and CD spectra peak optimization.

However, our results also suggest that the benefit of incorporating stereochemistry is less pronounced in tasks where other molecular features, such as size or functional group presence, may play a more dominant role. This was evident in the protein-ligand docking task for target 6Y2F, where the impact of stereochemistry was less substantial due to the larger and less constrained binding pocket. This observation underscores the need for a considered approach when deciding on the necessity of stereochemical information in generative models.

This work provides insights into the capabilities and limitations of current string-based generative models in capturing and leveraging stereochemical information for molecular design. While the incorporation of such information can be beneficial, particularly for applications where 3D molecular structure is critical, the decision should be guided by a careful assessment of the task-specific requirements and the trade-offs associated with increased model complexity. Further investigation into more efficient and robust representations of stereochemistry, for example through graph-based representations, is a direction for future research.

Data Availability

The code and data, including the implemented models and benchmarking tasks, are available at https://github.com/aspuru-guzik-group/stereogeneration under the MIT license.

Acknowledgements

The authors thank Austin Cheng, Cher Tian Ser, Leon Schlosser, and AkshatKumar Nigam for helpful discussions. G.T. acknowledges the support of the Vector Institute for Artificial Intelligence, and the Natural Sciences and Engineering Research Council of Canada (NSERC). K.J. acknowledges funding through an International Postdoc grant from the Swedish Research Council (no. 2020-00314). A.A.-G. acknowledges support from the Canada 150 Research Chairs program and CIFAR, as well as the generous support of Anders G. Frøseth. Computations were made on the supercomputers Béluga and Narval from École de technologie supérieure, managed by Calcul Québec and the Digital Research Alliance of Canada. The operation of the supercomputers is funded by the Canada Foundation for Innovation (CFI), Ministère de l'Économie, des Sciences et de l'Innovation du Québec (MESI) and le Fonds de recherche du Québec–Nature et technologies (FRQ-NT).

[1] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

[2] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[3] Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science*, 4(1):120–131, 2018.

[4] Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *arXiv preprint arXiv:1805.11973*, 2018.

[5] Joshua Meyers, Benedek Fabian, and Nathan Brown. De novo molecular design and generative models. *Drug Discovery Today*, 26(11):2707–2715, 2021.

[6] Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. *Advances in Neural Information Processing Systems*, 34:27381–27394, 2021.

[7] Zhenpeng Yao, Benjamín Sánchez-Lengeling, N Scott Bobbitt, Benjamin J Bucior, Sai Govind Hari Kumar, Sean P Collins, Thomas Burns, Tom K Woo, Omar K Farha, Randall Q Snurr, et al. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence*, 3(1):76–86, 2021.

[8] Francesca Grisoni, Berend JH Huisman, Alexander L Button, Michael Moret, Kenneth Atz, Daniel Merk, and Gisbert Schneider. Combining generative artificial intel-

ligence and on-chip synthesis for de novo drug design. *Science Advances*, 7(24):eabg3338, 2021.

[9] Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature Biotechnology*, 37(9):1038–1040, 2019.

[10] Feng Ren, Xiao Ding, Min Zheng, Mikhail Korzinkin, Xin Cai, Wei Zhu, Alexey Mantsyzov, Alex Aliper, Vladimir Aladinskiy, Zhongying Cao, et al. Alphafold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel cdk20 small molecule inhibitor. *Chemical Science*, 14(6):1443–1452, 2023.

[11] Gisbert Schneider and Uli Fechner. Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*, 4(8):649–663, 2005.

[12] Petra Schneider, W Patrick Walters, Alleyn T Plowright, Norman Sieroka, Jennifer Listgarten, Robert A Goodnow Jr, Jasmin Fisher, Johanna M Jansen, José S Duca, Thomas S Rush, et al. Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery*, 19(5):353–364, 2020.

[13] Robert Pollice, Gabriel dos Passos Gomes, Matteo Aldeghi, Riley J Hickman, Mario Krenn, Cyrille Lavigne, Michael Lindner-DAddario, AkshatKumar Nigam, Cher Tian Ser, Zhenpeng Yao, et al. Data-driven strategies for accelerated materials design. *Accounts of Chemical Research*, 54(4):849–860, 2021.

[14] Gary Tom, Stefan P Schmid, Sterling G Baird, Yang Cao, Kourosh Darvish, Han Hao, Stanley Lo, Sergio Pablo-García, Ella M Rajaonson, Marta Skreta, et al. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 2024.

[15] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International Conference on Machine Learning*, pages 1945–1954. PMLR, 2017.

[16] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International Conference on Machine Learning*, pages 2323–2332. PMLR, 2018.

[17] Robin Winter, Floriane Montanari, Frank Noé, and Djork-Arné Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical Science*, 10(6):1692–1701, 2019.

[18] Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in Pharmacology*, 11:565644, 2020.

[19] Gabriel Lima Guimaraes, Benjamin Sanchez-Lengeling, Carlos Outeiral, Pedro Luis Cunha Farias, and Alán Aspuru-Guzik. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843*, 2017.

[20] Benjamin Sanchez-Lengeling, Carlos Outeiral, Gabriel L Guimaraes, and Alan Aspuru-Guzik. Optimizing distributions over molecular space. an objective-reinforced generative adversarial network for inverse-design chemistry (organic). *ChemRxiv*, 2017.

[21] Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9:1–14, 2017.

[22] Thomas Blaschke, Josep Arús-Pous, Hongming Chen, Christian Margreitter, Christian Tyrchan, Ola Engkvist, Kostas Papadopoulos, and Atanas Patronov. Reinvent 2.0: an ai tool for de novo drug design. *Journal of Chemical Information and Modeling*, 60(12):5918–5922, 2020.

[23] Andrei Cristian Nica, Moksh Jain, Emmanuel Bengio, Cheng-Hao Liu, Maksym Korablyov, Michael M Bronstein, and Yoshua Bengio. Evaluating generalization in gflownets for molecule design. In *ICLR2022 Machine Learning for Drug Discovery*, 2022.

[24] Hannes H Loeffler, Jiazhen He, Alessandro Tibo, Jon Paul Janet, Alexey Voronov, Lewis H Mervin, and Ola Engkvist. Reinvent 4: Modern ai–driven generative molecule design. *Journal of Cheminformatics*, 16(1):20, 2024.

[25] Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific Reports*, 9(1):10752, 2019.

[26] Jeff Guo and Philippe Schwaller. Saturn: Sample-efficient generative molecular design using memory manipulation. *arXiv preprint arXiv:2405.17066*, 2024.

[27] Nathan Brown, Ben McKay, François Gilardoni, and Johann Gasteiger. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *Journal of Chemical Information and Computer Sciences*, 44(3):1079–1087, 2004.

[28] Jan H Jensen. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical Science*, 10(12):3567–3572, 2019.

[29] AkshatKumar Nigam, Robert Pollice, and Alán Aspuru-Guzik. Parallel tempered genetic algorithm guided by deep neural networks for inverse molecular design. *Digital Discovery*, 1(4):390–404, 2022.

[30] Austin Tripp and José Miguel Hernández-Lobato. Genetic algorithms are strong baselines for molecule generation. *arXiv preprint arXiv:2310.09267*, 2023.

[31] Orion Dollar, Nisarg Joshi, David AC Beck, and Jim Pfaendtner. Attention-based generative models for de novo molecular design. *Chemical Science*, 12(24):8362–8372, 2021.

[32] Viraj Bagal, Rishal Aggarwal, PK Vinod, and U Deva Priyakumar. Molgpt: molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076, 2021.

[33] Jannis Born and Matteo Manica. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nature Machine Intelligence*, 5(4):432–444, 2023.

[34] Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024.

[35] Yuanqi Du, Arian R Jamasb, Jeff Guo, Tianfan Fu, Charles Harris, Yingheng Wang, Chenru Duan, Pietro Liò, Philippe Schwaller, and Tom L Blundell. Machine learning-aided generative molecular design. *Nature Machine Intelligence*, pages 1–16, 2024.

[36] Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of Chemical Information and Modeling*, 58(9):1736–1741, 2018.

[37] Daniel Flam-Shepherd, Kevin Zhu, and Alán Aspuru-Guzik. Language models can learn complex molecular distributions. *Nature Communications*, 13(1):3293, 2022.

[38] Michael A Skinnider, R Greg Stacey, David S Wishart, and Leonard J Foster. Chemical language models enable navigation in sparsely populated chemical space. *Nature Machine Intelligence*, 3(9):759–770, 2021.

[39] Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of Chemical Information and Modeling*, 59(3):1096–1108, 2019.

[40] Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor Coley. Sample efficiency matters: a benchmark for practical molecular optimization. *Advances in Neural Information Processing Systems*, 35:21342–21357, 2022.

[41] Miguel García-Ortegón, Gregor NC Simm, Austin J Tripp, José Miguel Hernández-Lobato, Andreas Bender, and Sergio Bacallado. Dockstring: easy molecular docking yields better benchmarks for ligand design. *Journal of Chemical Information and Modeling*, 62(15):3486–3502, 2022.

[42] Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

[43] AkshatKumar Nigam, Robert Pollice, Gary Tom, Kjell Jorner, John Willes, Luca Thiede, Anshul Kundaje, and Alán Aspuru-Guzik. Tartarus: A benchmarking platform for realistic and practical inverse molecular design. *Advances in Neural Information Processing Systems*, 36:3263–3306, 2023.

[44] Scott A Wildman and Gordon M Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, 1999.

[45] G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2012.

[46] BenevolentAI. BenevolentAI/guacamol_results. `https://github.com/BenevolentAI/guacamol_results`, 2020. Accessed on August 19, 2024.

[47] Sungsoo Ahn, Junsu Kim, Hankook Lee, and Jinwoo Shin. Guiding deep molecular optimization with genetic exploration. *Advances in Neural Information Processing Systems*, 33:12008–12021, 2020.

[48] Austin Tripp, Gregor NC Simm, and José Miguel Hernández-Lobato. A fresh look at de novo molecular design benchmarks. In *NeurIPS 2021 AI for Science Workshop*, 2021.

[49] Philipp Renz, Dries Van Rompaey, Jörg Kurt Wegner, Sepp Hochreiter, and Günter Klambauer. On failure modes in molecule generation and optimization. *Drug Discovery Today: Technologies*, 32:55–63, 2019.

[50] Philippe Gendreau, Joseph-André Turk, Nicolas Drizard, Vinicius Barros Ribeiro da Silva, Clarisse Descamps, and Yann Gaston-Mathé. Molecular assays simulator to unravel predictors hacking in goal-directed molecular generations. *Journal of Chemical Information and Modeling*, 63(13):3983–3998, 2023.

[51] Jiankun Lyu, John J Irwin, and Brian K Shoichet. Modeling the expansion of virtual screening libraries. *Nature Chemical Biology*, 19(6):712–718, 2023.

[52] Andreas Bender, Nadine Schneider, Marwin Segler, W Patrick Walters, Ola Engkvist, and Tiago Rodrigues. Evaluation guidelines for machine learning tools in the chemical sciences. *Nature Reviews Chemistry*, 6(6):428–442, 2022.

[53] Naveen Chhabra, Madan L Aseri, and Deepak Padmanabhan. A review of drug isomerism and its significance. *International Journal of Applied and Basic Medical Research*, 3(1):16–18, 2013.

[54] Jonathan McConathy and Michael J Owens. Stereochemistry in drug action. *Primary care companion to the Journal of Clinical Psychiatry*, 5(2):70, 2003.

[55] Rebecca U McVicker and Niamh M OBoyle. Chirality of new drug approvals (2013–2022): trends and perspectives. *Journal of Medicinal Chemistry*, 67(4):2305–2320, 2024.

[56] Silas W Smith. Chiral toxicology: it's the same thing only different. *Toxicological Sciences*, 110(1):4–30, 2009.

[57] Stephan Guy, Laure Guy, Amina Bensalah-Ledoux, Antonio Pereira, Vincent Grenard, Olivier Cosso, and Teophile Vautey. Pure chiral organic thin films with high isotropic optical activity synthesized by uv pulsed laser deposition. *Journal of Materials Chemistry*, 19(38):7093–7097, 2009.

[58] Erick M Carreira and Lisbet Kvaerno. *Classics in stereoselective synthesis*. John Wiley & Sons, 2009.

[59] Gianluigi Albano, Gennaro Pescitelli, and Lorenzo Di Bari. Chiroptical properties in thin films of π-conjugated systems. *Chemical Reviews*, 120(18):10145–10243, 2020.

[60] Audrey Cuvellier, Robrecht Verhelle, Joost Brancart, Bram Vanderborght, Guy Van Assche, and Hubert Rahier. The influence of stereochemistry on the reactivity of the diels–alder cycloaddition and the implications for reversible network polymerization. *Polymer Chemistry*, 10(4):473–485, 2019.

[61] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[62] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

[63] Austin H Cheng, Andy Cai, Santiago Miret, Gustavo Malkomes, Mariano Phielipp, and Alán Aspuru-Guzik. Group selfies: a robust fragment-based molecular string representation. *Digital Discovery*, 2(3):748–758, 2023.

[64] Mariami Basilaia, Matthew H Chen, Jim Secka, and Jeffrey L Gustafson. Atropisomerism in the pharmaceutically relevant realm. *Accounts of Chemical Research*, 55(20):2904–2919, 2022.

[65] Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, et al. Selfies and the future of molecular string representations. *Patterns*, 3(10), 2022.

[66] Fabrizio Mastrolorito, Fulvio Ciriaco, Maria Vittoria Togo, Nicola Gambacorta, Daniela Trisciuzzi, Cosimo Damiano Altomare, Nicola Amoroso, Francesca Grisoni, and Orazio Nicolotti. fragsmiles: A chemical string notation for advanced fragment and chirality representation. 2024.

[67] Teague Sterling and John J Irwin. Zinc 15–ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015.

[68] Greg Landrum. RDKit: Open-source cheminformatics (v2024.03.5). `https://www.rdkit.org`, 2024.

[69] AkshatKumar Nigam, Robert Pollice, Mario Krenn, Gabriel dos Passos Gomes, and Alan Aspuru-Guzik. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (stoned) algorithm for molecules using selfies. *Chemical Science*, 12(20):7079–7090, 2021.

[70] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

[71] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[72] David Ryan Koes, Matthew P Baumgartner, and Carlos J Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, 2013.

[73] Thomas A Halgren. MMFF vi. MMFF94s option for energy minimization studies. *Journal of Computational Chemistry*, 20(7):720–729, 1999.

[74] Jacob A Olsen, David W Banner, Paul Seiler, Ulrike Obst Sander, Allan D'Arcy, Martine Stihle, Klaus Müller, and François Diederich. A fluorine scan of thrombin inhibitors to map the fluorophilicity/fluorophobicity of an enzyme active site: Evidence for C–F$\cdots$C=O interactions. *Angewandte Chemie International Edition*, 42(22):2507–2511, 2003.

[75] Norma J Greenfield. Using circular dichroism spectra to estimate protein secondary structure. *Nature Protocols*, 1(6):2876–2890, 2006.

[76] Warren N Herman. Polarization eccentricity of the transverse field for modes in chiral core planar waveguides. *JOSA A*, 18(11):2806–2818, 2001.

[77] Ezekiel Bahar. Mueller matrices for waves reflected and transmitted through chiral materials: waveguide modal solutions and applications. *JOSA B*, 24(7):1610–1619, 2007.

[78] Peter Lodahl, Sahand Mahmoodian, Søren Stobbe, Arno Rauschenbeutel, Philipp Schneeweiss, Jürgen Volz, Hannes Pichler, and Peter Zoller. Chiral quantum optics. *Nature*, 541(7638):473–480, 2017.

[79] Stefan Grimme. Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations. *Journal of Chemical Theory and Computation*, 15(5):2847–2862, 2019.

[80] Philipp Pracht, Fabian Bohle, and Stefan Grimme. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics*, 22(14):7169–7192, 2020.

[81] Philipp Pracht, Stefan Grimme, Christoph Bannwarth, Fabian Bohle, Sebastian Ehlert, Gereon Feldmann, Johannes Gorges, Marcel Müller, Tim Neudecker, Christoph Plett, et al. Cresta program for the exploration of low-energy molecular chemical space. *The Journal of Chemical Physics*, 160(11), 2024.

[82] Christoph Bannwarth, Eike Caldeweyher, Sebastian Ehlert, Andreas Hansen, Philipp Pracht, Jakob Seibert, Sebastian Spicher, and Stefan Grimme. Extended tight-binding quantum chemistry methods. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 11(2):e1493, 2021.

[83] Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtban accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of Chemical Theory and Computation*, 15(3):1652–1671, 2019.

[84] Christoph Bannwarth and Stefan Grimme. A simplified time-dependent density functional theory approach for electronic ultraviolet and circular dichroism spectra of very large molecules. *Computational and Theoretical Chemistry*, 1040:45–53, 2014.

[85] Stefan Grimme and Christoph Bannwarth. Ultra-fast computation of electronic spectra for large systems by tight-binding based simplified tamm-dancoff approximation (stda-xtb). *The Journal of Chemical Physics*, 145(5), 2016.

Supplementary Information:
Stereochemistry-aware string-based molecular generation

A. Fitness functions

The targets of the rediscovery tasks are shown in Figure S1. For the docking tasks, the native ligands for the protein targets are shown in Figure S2. Samples of CD spectra generated using `sTDA-xTB` for some example stereoisomers are shown in Figure S3.

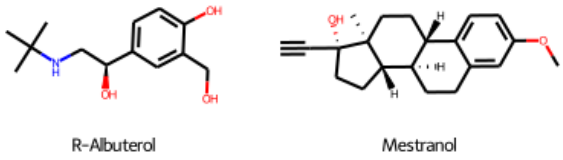


FIG. S1. **Rediscovery targets.** The chemical structures of the rediscovery targets (R)-albuterol and mestranol. (R)-albuterol has a single chiral centre, while mestranol has five.

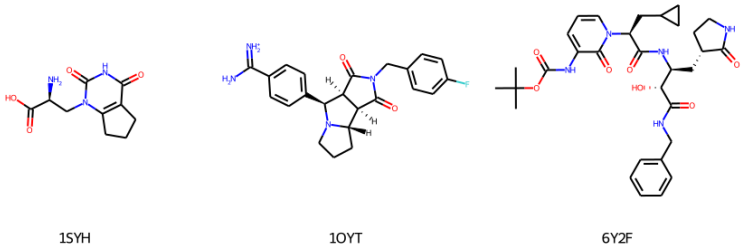


FIG. S2. **Structure of native ligands.** The structures of the native ligands of the proteins.

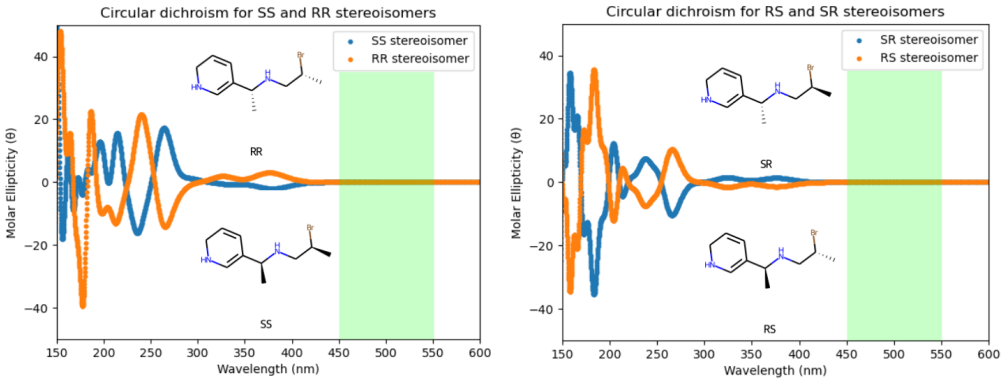


FIG. S3. **Examples of CD spectra.** Example of CD spectra generated from `sTDA-xTB`. Mirror opposite stereoisomers produce spectra that are inverted with respect to each other. Highlighted region is the region of interest for the peak score.

B. Additional results

A table of maximum achieved scores, analogous to Table I, is found in Table S2. The best molecules generated by each model for each task are found in the Figures S4 to S9.

		REINVENT	JANUS	GroupJANUS
Non-stereo	(R)-albuterol rediscovery	0.796 ± 0.184	0.882 ± 0.124	0.905 ± 0.122
	Mestranol rediscovery	0.349 ± 0.050	0.683 ± 0.043	0.703 ± 0.042
	1SYH docking	10.900 ± 0.377	12.170 ± 0.615	12.850 ± 0.792
	1OYT docking	11.960 ± 0.246	13.480 ± 0.565	13.350 ± 0.528
	6Y2F docking	9.780 ± 0.244	10.350 ± 0.519	10.510 ± 0.357
	CD spectral peak score	1622 ± 432	6883 ± 1700	7043 ± 2937
Stereo	(R)-albuterol rediscovery	0.730 ± 0.201	**1.000 ± 0.000**	**1.000 ± 0.000**
	Mestranol rediscovery	0.339 ± 0.045	**0.925 ± 0.121**	**0.982 ± 0.056**
	1SYH docking	10.810 ± 0.423	12.340 ± 0.564	13.040 ± 1.041
	1OYT docking	11.770 ± 0.309	**14.040 ± 0.477**	13.470 ± 0.677
	6Y2F docking	9.870 ± 0.497	10.710 ± 0.621	10.580 ± 0.439
	CD spectral peak score	1400 ± 390	**9533 ± 3351**	6845 ± 1931

TABLE S2. **Maximum achieved score for all tasks, for stereo and non-stereo aware models.** The mean and standard deviation are reported. Statistically significantly higher scores between the non-stereo and stereo variants are bolded.

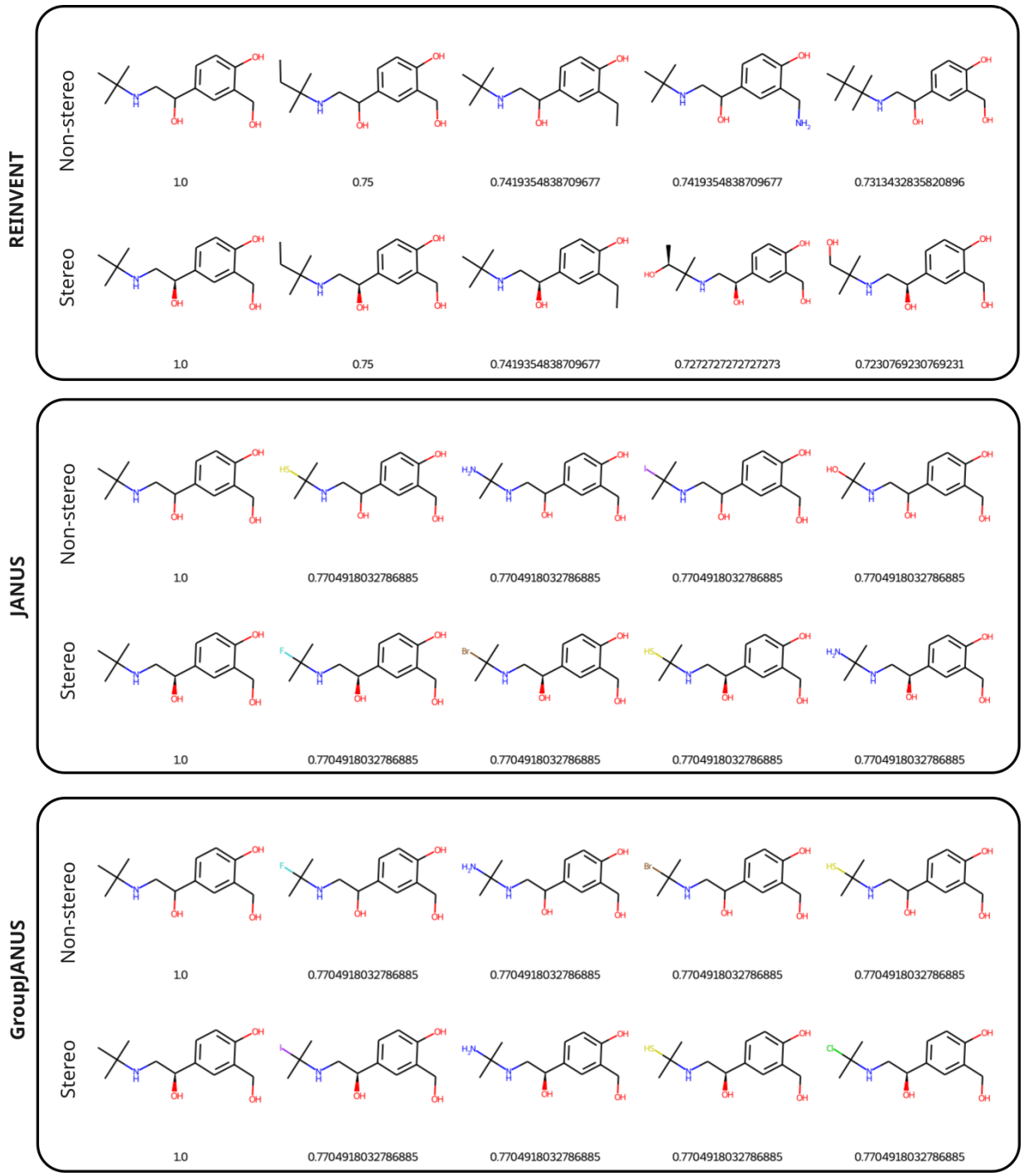


FIG. S4. **Top molecules found for (R)-albuterol rediscovery.** Top 5 compounds across all runs on (R)-albuterol rediscovery for each the non-stereo and stereo versions of `REINVENT`, `JANUS`, and `GroupJANUS`. Only the best of duplicates are retained.

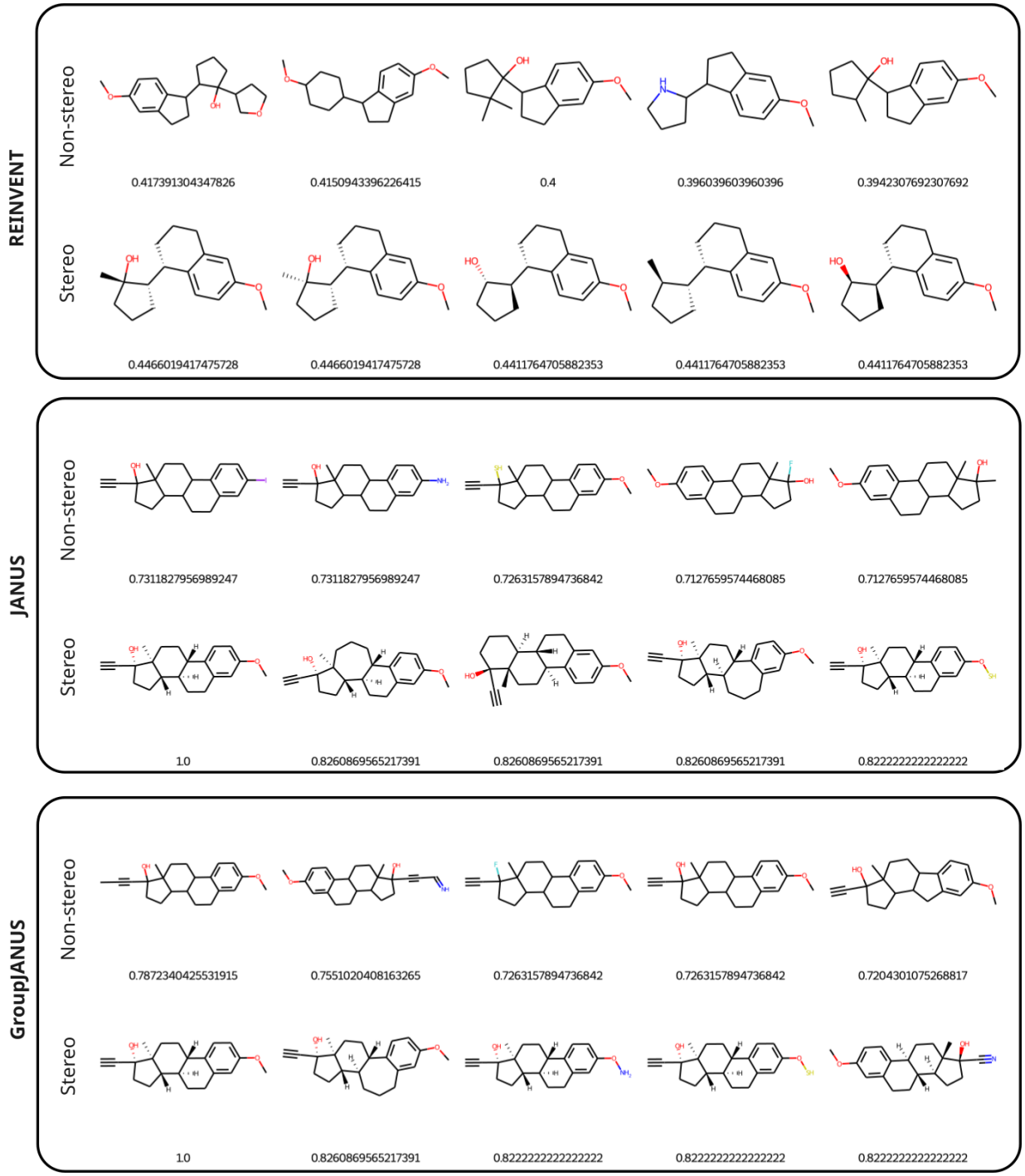


FIG. S5. **Top molecules found for mestranol rediscovery.** Top 5 compounds across all runs on mestranol rediscovery for each the non-stereo and stereo versions of `REINVENT`, `JANUS`, and `GroupJANUS`. Only the best of duplicates are retained.

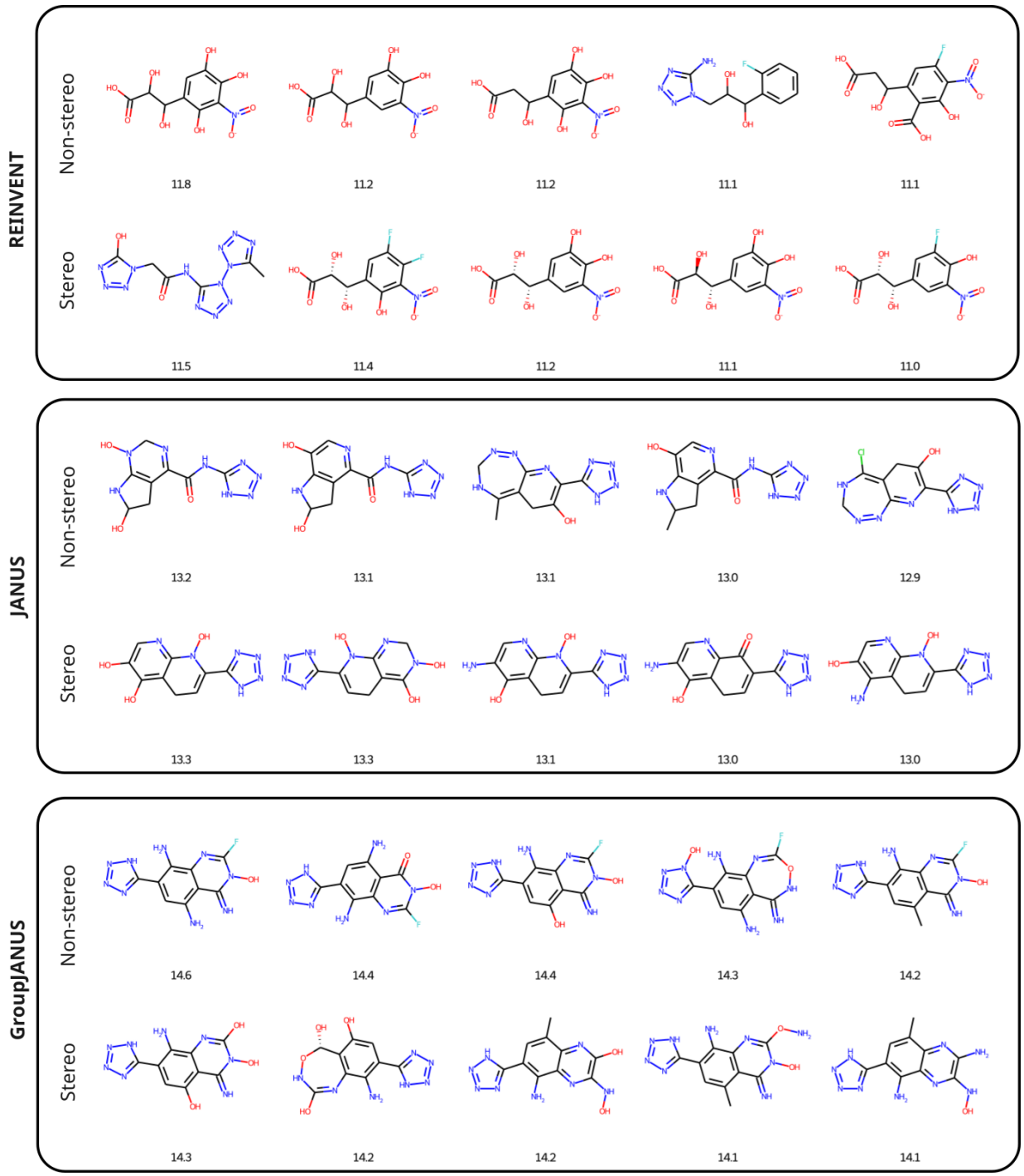


FIG. S6. **Top molecules found for 1SYH docking task.** Top 5 compounds across all runs on 1SYH docking for each the non-stereo and stereo versions of `REINVENT`, `JANUS`, and `GroupJANUS`. Only the best of duplicates are retained.

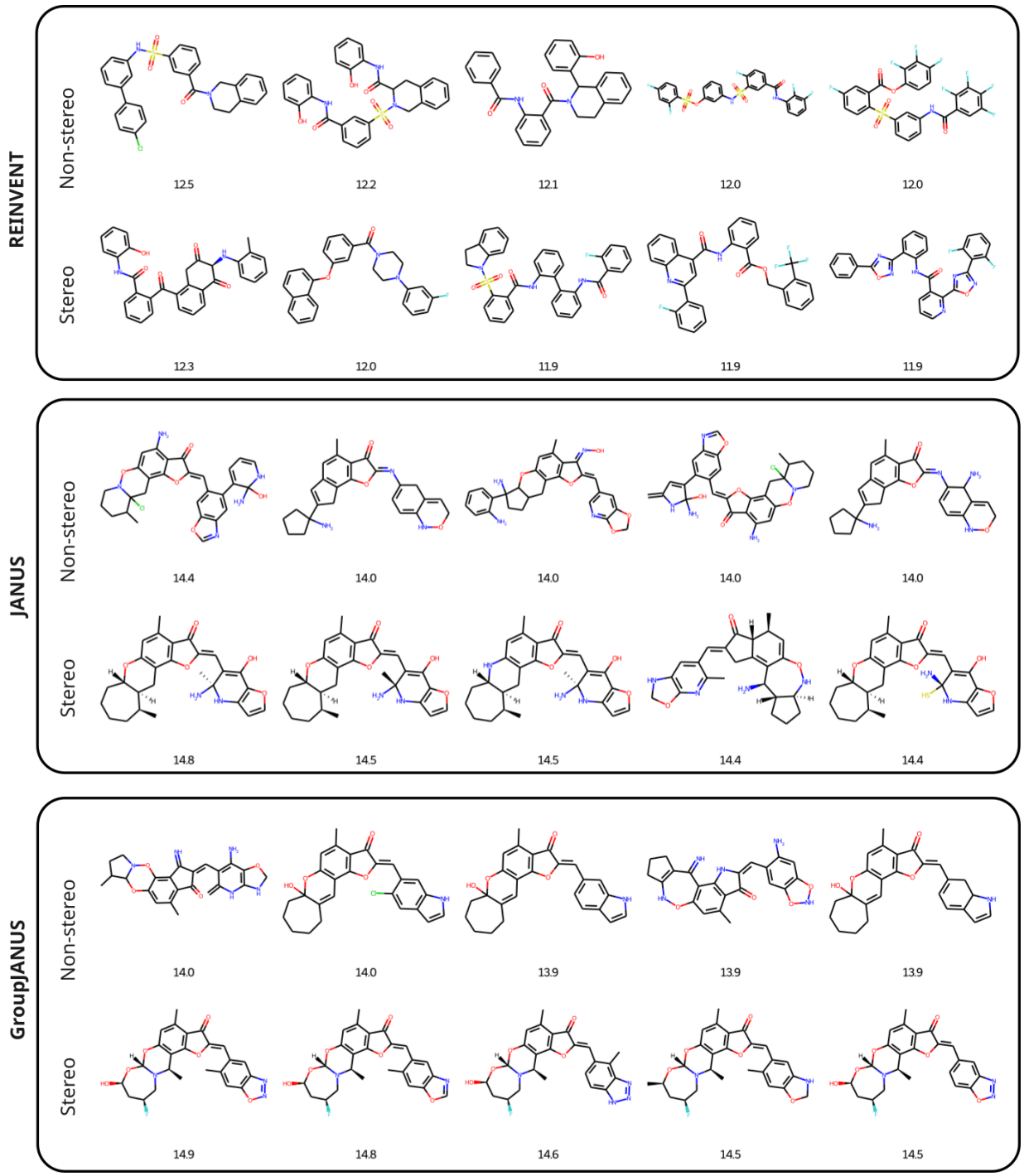


FIG. S7. **Top molecules found for 1OYT docking task.** Top 5 compounds across all runs on 1OYT docking for each the non-stereo and stereo versions of `REINVENT`, `JANUS`, and `GroupJANUS`. Only the best of duplicates are retained.

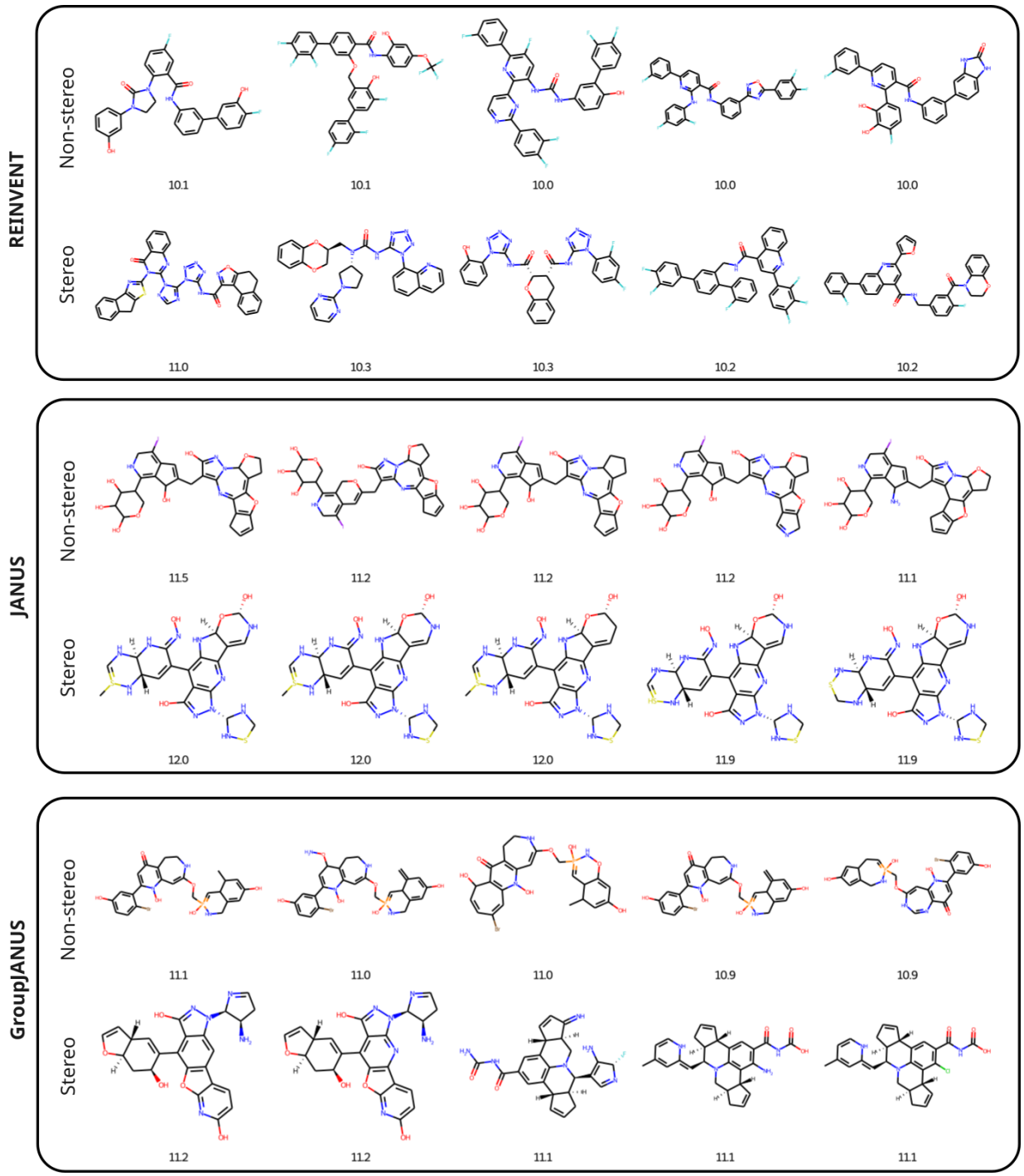


FIG. S8. **Top molecules found for 6Y2F docking task.** Top 5 compounds across all runs on 6Y2F docking for each the non-stereo and stereo versions of `REINVENT`, `JANUS`, and `GroupJANUS`. Only the best of duplicates are retained.

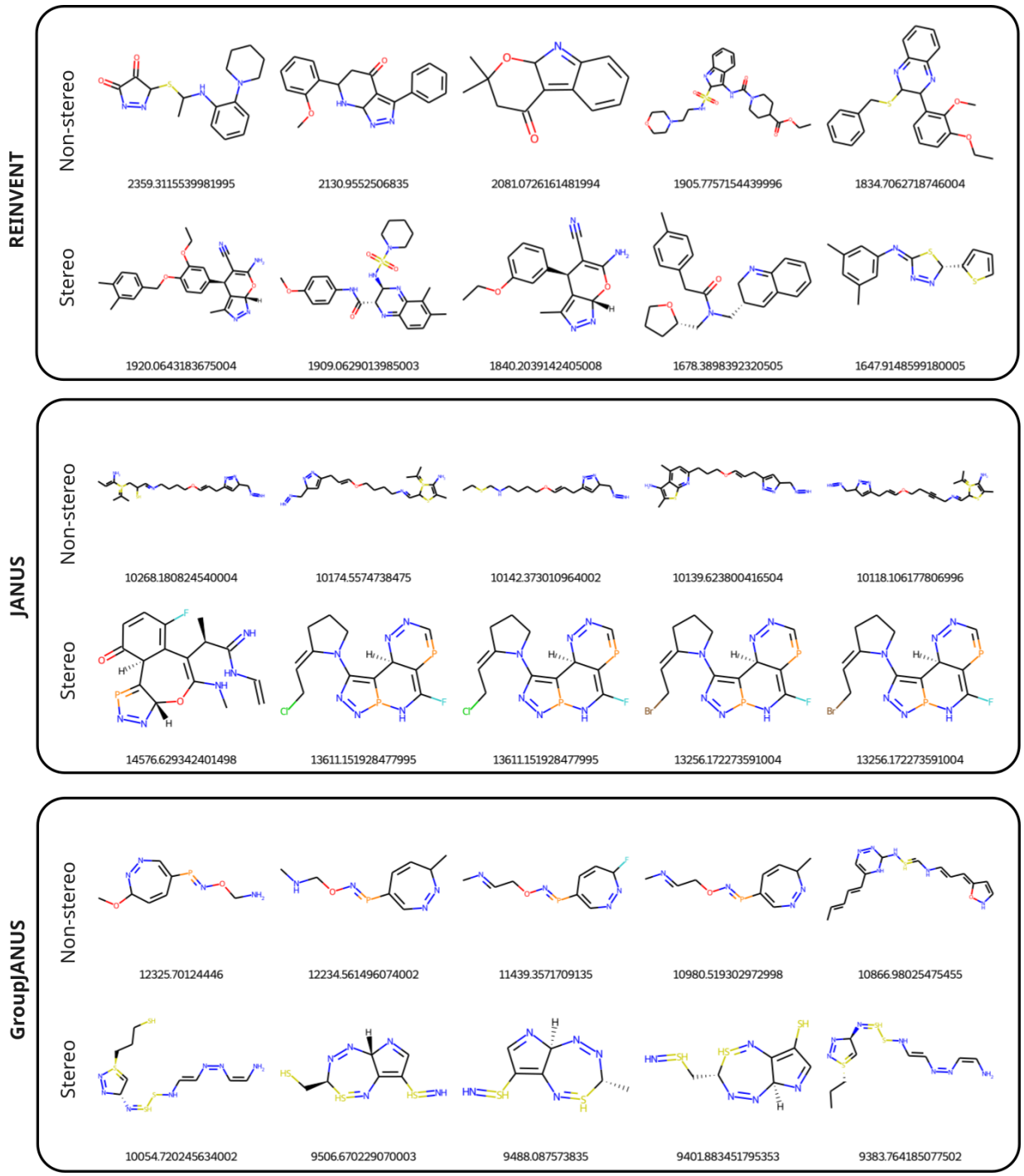


FIG. S9. **Top molecules found for CD spectra task.** Top 5 compounds across all runs on CD spectra task for each the non-stereo and stereo versions of `REINVENT`, `JANUS`, and `GroupJANUS`. Only the best of duplicates are retained.