Decoupled coordinates for machine learning-based molecular fragment linking

MACHINE
LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
17 November 2021

REVISED
17 January 2022

ACCEPTED FOR PUBLICATION
1 February 2022

PUBLISHED
22 February 2022

PAPER

Decoupled coordinates for machine learning-based molecular fragment linking

Markus Fleck, Michael Müller, Noah Weber and Christopher Trummer*

Celeris Therapeutics GmbH, 8010 Graz, Austria

* Author to whom any correspondence should be addressed.

E-mail: c.trummer@celeristx.com

Keywords: fragment linking, deep generative models, drug design

Abstract

Recent developments in machine learning-based molecular fragment linking have demonstrated the importance of informing the generation process with structural information specifying the relative orientation of the fragments to be linked. However, such structural information has so far not been provided in the form of a complete relative coordinate system. We present a decoupled coordinate system consisting of bond lengths, bond angles and torsion angles, and show that it is complete. By incorporating this set of coordinates in a linker generation framework, we show that it has a significant impact on the quality of the generated linkers. To elucidate the advantages of such a coordinate system, we investigate the amount of reliable information within the different types of degrees of freedom using both detailed ablation studies and an information-theoretical analysis. The presented benefits suggest the application of a complete and decoupled relative coordinate system as a standard good practice in linker design.

1. Introduction

Computational drug design remains a challenging problem, primarily due to the vast size of the drug-like chemical space [1–4]. A sub-problem of molecular generation is the task of fragment linking: given a pair of structures, the goal of the design process is their connection via an appropriate linker. This process of generating larger molecules from pre-determined fragments is central to current targeted protein degradation approaches, which have become a major focus of structure-based drug design in recent years (see [5] for a recent review).

For targeted protein degradation, a fragment binding to an E3 ligase and a fragment binding to a target protein of interest need to be joined by a linker. It is well-known that geometric considerations are crucial for successful linker design in this context. For example, the length of the linker between the two binding fragments plays an important role for a compound’s efficacy as a drug [6, 7], with linker lengths below some threshold leading to ineffective pharmaceuticals [8]. Accordingly, the linker design process should take into account the desired distance between the fragments. In addition, the resulting linker should avoid interfering with the binding modes of the individual fragments so that the resulting compound maintains high activity [9, 10]. Thus, an effective linker generation method should incorporate structural information such as the relative distance and orientation between the molecules to be joined.

Recently, a linker generation method using such structural information, called DeLinker [11], has been proposed. DeLinker is based on a machine learning framework operating on molecular graphs and provides a generative model allowing to create different linkers for a given pair of fragments. While the results of the method arguably constituted a breakthrough in machine learning-based fragment linking, the structural information employed by the method is problematic as it consists only of the fragments’ distance and angle.

Conversely, the benefits of a complete decoupled coordinate system are well known in computational chemistry and biophysics. With the information processing nature of machine learning models in mind, a relevant example here is the calculation of configurational entropy [12–19] from molecular mechanics simulations. Due to physico-chemical forces, the molecular coordinates are subject to soft constraints. For instance, the distance between two neighboring hydrogen atoms in a methyl group hardly varies. Such soft constraints narrow the accessible phase space of the molecule and therefore lower the entropy. In Cartesian coordinates, these soft constraints need to be accommodated by adequate sampling, which quickly becomes prohibitive with increasing molecular size. Another example involves the mathematically rigorous separation of the placeholder (dummy) atom partition function from the physical partition function in alchemical free energy simulations [20]. Here, if the transformed molecule comprises less atoms then the original molecule, dummy atoms are utilized. The dummy atoms need to be attached to the physical part of the transformed molecule by applying forces only to a selected set of decoupled coordinates. These coordinates are selected for the absence of geometrical constraints between the physical and the dummy coordinates (otherwise, the physical partition function is coupled to the dummy partition function, which leads to inaccuracies). A simple example for a geometrical constraint would be the fact that three angles in a triangle sum up to 180 degrees, so only two angles can be chosen freely. Then, e.g. if two (force-carrying) angles reside in the physical molecule and one is part of the dummy atom group, the physical partition function does not factorize from the dummy atom partition function and systematic bias arises. Note that these geometrical constraints are hard mathematical constraints. They differ from the soft constraints mentioned above, which stem from physico-chemical forces rather than from pure geometrical aspects.

Based on bond lengths, bond angles and torsion angles, coordinate systems which are by definition free from geometrical constraints can be formulated. Importantly, such coordinate systems feature physico-chemical decoupling naturally by accommodating the molecular topology made up from rigid bonds. In this work, we apply such a bond-angle-torsion (BAT) coordinate system in order to specify the relative orientation of molecular fragments to be linked. As opposed to the coordinates applied in previous work [11], the coordinate system proposed in this article is mathematically well-defined, complete and decoupled. We demonstrate the advantage of such a coordinate system by incorporating it into the DeLinker framework [11] and performing detailed analyses regarding the information content in the different degrees of freedom. Our results highlight the advantage of informing linker generation methods by a complete and decoupled set of coordinates over the current practice of using partial structural information.

2. Methods

In this article, we propose a beneficial choice of a relative coordinate system for linking pairs of molecular fragments by specifying their relative orientation in 3D space. In order to benchmark our proposed coordinate system, the details of which will be given below, we follow DeLinker [11], a recent pioneering machine learning framework for linking molecular fragments.

2.1. Overall model framework

Here, we give a brief summary of the DeLinker [11] framework. For further details, we refer to [11] and the references therein (in particular [21–23]).

From a high-level point of view, DeLinker [11] is embedded in a variational autoencoder (VAE) [24] framework. Linker generation is performed by seeding the latent variables of the decoder of the VAE using two fragments as input. First, the graph representation of the fragment pair is encoded by a standard gated graph neural network [23]. Then, a set of atoms is initialized in order to serve as expansion nodes for linking a pair of fragments. The maximum linker length is given by the chosen size of this set of expansion nodes. For each atom, a multidimensional hidden state is drawn from standard normal distributions. The atom type of the respective node is derived via sampling from a learned mapping from the hidden state to a Boltzmann distribution (i.e. softmax or, more precisely, softargmax).

Successively, the fragments are linked by attaching atoms from the set of expansion nodes in an iterative manner. Following the breadth-first paradigm, a first-in-first-out queue is initialized with two exit atoms (figure 1), one per fragment. The focused node, i.e. the first node in the queue, forms covalent bonds to candidate atoms, which are themselves added to the queue upon bond formation. The selection process of the focused node is terminated upon forming a bond to a special stop node. Then, the focused node is removed from the queue and the next atom in turn initiates bond formation. Nodes become closed once they are focused, which means they are no longer considered as candidate nodes for bond formation throughout the entire generation process. The bond selection procedure is accomplished by computing feature vectors between the focused atom and all candidate atoms. These feature vectors comprise both atom types, their hidden states as well as their graph distance. Furthermore, the bond formation process takes as an input the average of the hidden states across all nodes during node initialization (i.e. iteration zero) as well as at the current iteration. The current iteration number is supplied in addition. Importantly, the feature vector is augmented with coordinates specifying the relative orientation between the fragments, the details of which will be given below. The actual chosen bond and its type (single, double or triple) are sampled from

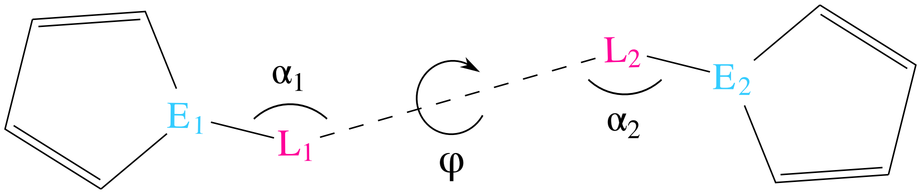

Figure 1. Bond-angle-torsion (BAT) coordinates. For specifying the relative orientation of two molecular fragments, we first define the exit atom E_1 of the first fragment as the atom which is connected to the initial linker atom L_1. E_2 and L_2 are defined analogously for the second fragment. Two of the BAT coordinates are constituted by the exit vectors $|\vec{b}_1| = |\vec{r}_{E_1 \to L_1}|$ and $|\vec{b}_3| = |\vec{r}_{L_2 \to E_2}|$, bridging the exit atoms to the initial linker atoms. A third BAT coordinate is defined by the length of the pseudo-bond $|\vec{b}_2| = |\vec{r}_{L_1 \to L_2}|$. The angles between $\vec{r}_{E_1 \to L_1}$ and $\vec{r}_{L_1 \to L_2}$ as well as between $\vec{r}_{L_1 \to L_2}$ and $\vec{r}_{L_2 \to E_2}$, termed α_1 and α_2 respectively, yield another two BAT coordinates. In order to obtain six BAT coordinates, a well-known number for specifying relative molecular orientations, we add the dihedral angle ϕ of the atom chain E_1–L_1–L_2–E_2.

Boltzmann distributions, which are based on learned mappings from the feature vectors and are masked with valency constraints [21].

After each bond formation, the hidden states of all nodes are updated with respect to the newly formed graph topology. This update is performed via a standard gated graph neural network [23], taking as input the initial states of all nodes. Starting from the states of the nodes at iteration zero, instead of the current state, prevents the network from learning assembly pathways [21].

The generative process ends when the first-in-first-out queue is empty. Note that the described procedure does not prevent the generation of unlinked fragments, which, however, constitutes the only mechanism leading to invalid molecules as an outcome.

Fragment linking constitutes a multimodal problem. Two fragments can be linked in many different ways. Inspired by Jin *et al* [22], a low-dimensional latent vector z is introduced in order for the model to accommodate this one-to-many mapping. To avoid difficulties known from computer vision [25], during training, z is derived from the encoding of the ground-truth linked fragments. In this manner, the model is encouraged to pay attention to the latent vector. Furthermore, z is regularized via the KL-divergence to the standard normal distribution.

2.2. Relative coordinates

In this section, we derive our proposed relative coordinate system for linking molecular fragments. The following derivation follows the bond-angle-torsion [12, 26–31] coordinate formalism, which closely relates to the z-matrix representation [31, 32].

We use the following nomenclature (see figure 1). The atoms of the fragments to which the linker is covalently bound are referred to as exit atoms; E_1 and E_2 for the two fragments, respectively. The linker atoms L_1 and L_2 are attached to E_1 and E_2 via a rotable bond. Naturally for the problem of linking molecular fragments, we are not concerned with the global position and orientation of the fragment pair as a whole. Rather, we are interested in the position and orientation of the fragments relative to each other. Thus, without loss of generality, exit atom 1 (E_1) is positioned at the origin of the Cartesian coordinate system, linker atom 1 (L_1) on the x-axis and linker atom 2 (L_2) in the x–y plane. Then, these anchored Cartesian coordinates [26, 27] are given as follows:

$$\begin{aligned}
\vec{r}_{E_1}^{\,\mathsf{T}} &= (0,0,0) \\
\vec{r}_{L_1}^{\,\mathsf{T}} &= (L_1^x,0,0) \\
\vec{r}_{L_2}^{\,\mathsf{T}} &= (L_2^x,L_2^y,0) \\
\vec{r}_{E_2}^{\,\mathsf{T}} &= (E_2^x,E_2^y,E_2^z).
\end{aligned} \tag{1}$$

In order to achieve ease of notation, we constitute the following definitions.

$$\begin{aligned}
\vec{b}_1^{\,\mathsf{T}} &\equiv \vec{r}_{E_1 \to L_1} \equiv \vec{r}_{L_1} - \vec{r}_{E_1} = (L_1^x,0,0) \\
\vec{b}_2^{\,\mathsf{T}} &\equiv \vec{r}_{L_1 \to L_2} \equiv \vec{r}_{L_2} - \vec{r}_{L_1} = (L_2^x - L_1^x, L_2^y, 0) \\
\vec{b}_3^{\,\mathsf{T}} &\equiv \vec{r}_{L_2 \to E_2} \equiv \vec{r}_{E_2} - \vec{r}_{L_2} = (E_2^x - L_2^x, E_2^y - L_2^y, E_2^z).
\end{aligned} \tag{2}$$

Then, our implemented BAT coordinates are given as:

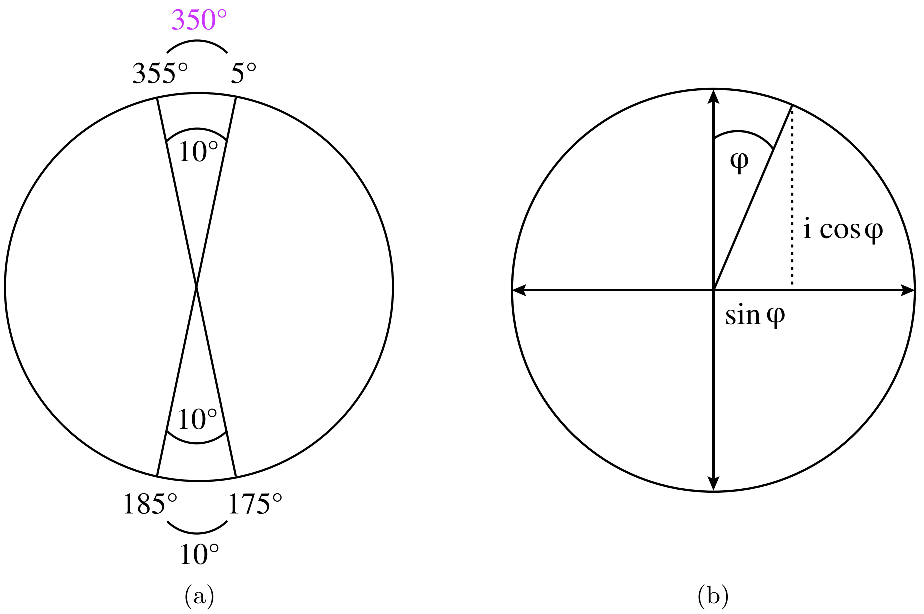

Figure 2. Treating periodicity. (a) shows an illustration of the problematic distance metric. Both the top and bottom angles measure 10 degrees geometrically. For the bottom angle, this is correctly reflected in its numerical value, i.e. $|\Delta\alpha_{\text{bottom}}| = |185° - 175°| = 10°$. For the top angle, however, the numerical value deviates from the underlying geometry due to the periodic discontinuity at $0° \mathrel{\hat{=}} 360°$. Here, we have $|\Delta\alpha_{\text{top}}| = |5° - 355°| = 350°$. This discontinuity of the mapping from geometric angles to their numerical values poses a challenge for machine learning algorithms. (b) The periodic discontinuity can be avoided by considering the dihedral an angle in the complex plane. Supplying both the real and imaginary part, i e. $\sin(\phi)$ and $\cos(\phi)$, respectively, dihedral angles are presented to the model in continuous form and without loss of information.

$$
\begin{aligned}
|\vec{b_1}| &= |L_1^x| \\
|\vec{b_2}| &= \sqrt{(L_2^x - L_1^x)^2 + (L_2^y)^2} \\
|\vec{b_3}| &= \sqrt{(E_2^x - L_2^x)^2 + (E_2^y - L_2^y)^2 + (E_2^z)^2} \\
\alpha_1 &= \arccos\left(\frac{\vec{b_1}^{\mathsf{T}} \cdot \vec{b_2}}{|\vec{b_1}||\vec{b_2}|}\right) \\
\alpha_2 &= \arccos\left(\frac{\vec{b_2}^{\mathsf{T}} \cdot \vec{b_3}}{|\vec{b_2}||\vec{b_3}|}\right) \\
\phi &= \arctan 2\left(\frac{|\vec{b_2}|[\vec{b_1}^{\mathsf{T}} \cdot (\vec{b_2} \times \vec{b_3})]}{|\vec{b_1} \times \vec{b_2}||\vec{b_2} \times \vec{b_3}|}, \frac{(\vec{b_1} \times \vec{b_2})^{\mathsf{T}} \cdot (\vec{b_2} \times \vec{b_3})}{|\vec{b_1} \times \vec{b_2}||\vec{b_2} \times \vec{b_3}|}\right).
\end{aligned}
\tag{3}
$$

Matrix multiplications and outer vector products are denoted by · and ×, respectively. The back-transformation from BAT (equation (3)) to anchored Cartesian coordinates [26, 27] (equation (1)), demonstrating the completeness of our proposed BAT coordinate system, is given in the appendix. In this context, note that we do not explicitly feed the model the lengths of the exit vectors, i.e. $|\vec{b_1}| = |\vec{r}_{E_1 \to L_1}|$ and $|\vec{b_3}| = |\vec{r}_{L_2 \to E_2}|$. As physical bond lengths, their values hardly vary in comparison to the pseudo-bond length $|\vec{b_2}| = |\vec{r}_{L_1 \to L_2}|$, which dominates the geometry. Thus, they barely carry any useful information for the model.

In contrast to bond angles, i.e. α_1 and α_2 in the present case, dihedral angles are periodic. This means $\alpha_1, \alpha_2 \in [0, \pi]$, but $\phi \in [0, 2\pi)$. Periodicity poses a challenge for machine learning algorithms, since the implied distance metric does not correctly reflect the underlying physical geometry. As done commonly (see e.g. [33–35]), we feed the model the sine and cosine of ϕ, instead of the plain dihedral angle value. One way to motivate this treatment states as follows. There is a periodic discontinuity (see figure 2(a)) at $0° \mathrel{\hat{=}} 360°$ in the mapping of the underlying geometry to its numerical dihedral angle value. A solution is inspired by the complex unit circle (figure 2(b)): One maps the dihedral ϕ to $\exp[i(\pi/2 - \phi)] = \sin(\phi) + i\cos(\phi)$. The machine learning model is fed the real and imaginary part of this transformation, i.e. both $\sin(\phi)$ and $\cos(\phi)$. These angular functions, as a representation of the periodic dihedral angle ϕ, are both continuous as well as faithful with respect to the underlying geometry. Additionally, they naturally map to the interval $[-1, 1]$, which constitutes another desirable property. Note that the transformation $\exp[i(\pi/2 - \phi)]$ was chosen over just $\exp(i\phi)$ in order to match the complex unit circle with the canonical definition of dihedral angles (figure 2(b)). However, as the order of the input features is irrelevant for the machine-learning algorithm, this choice of swapping sine with cosine is arbitrary.

2.3. Data preparation

Following DeLinker [11], we used two datasets, a selected [36] set of 250 000 ZINC [37] compounds as well as CASF-2016 (i.e. the PDBbind core set) [38]. A major difference between these two sets is the origin of the 3D structural information. While for CASF, the structures stem from 285 high-quality crystal structures of ligands bound to proteins, the ZINC structures were generated *in silico* using RDKit [39], as detailed below.

For both data sets, preparation was performed in the same manner. The ligands were split into three parts, i.e. two fragments and the corresponding linker. Cuts were performed on acyclic single bonds outside of functional groups [40]. Triplet splits for which either of the three components is unrealistically small, or the linker is inflated with respect to the fragments, were removed. The remaining triplet splits were filtered further by molecular graph (2D) properties, in particular synthetic accessibility (SA) [41], ring aromaticity as well as pan-assay interference compounds (PAINS) [42]. This procedure yielded ∼420 000 triplet splits for the ZINC data set and 309 triplet splits for CASF.

Training was performed purely on the ZINC data set. 400 compounds each were randomly selected from the whole ZINC set and held back as a validation and test set. The CASF data was used solely as a test set for evaluation.

As outlined above, unlike for CASF, where experimental 3D structures are available, the ZINC 3D coordinates needed to be generated. This was accomplished using RDKit [39] via employing the Merck molecular force field (MMFF) [43, 44]. The DeLinker [11] code without the 3D structures of the training set (most likely due to storage space reasons). For calculating our proposed relative coordinates of the training set, we reproduced the 3D structures. DeLinker [11] uses the distance from E_1 to E_2 as well as the angle between the exit vectors (i.e. $\vec{r}_{E_1\rightarrow L_1}$ and $\vec{r}_{E_2\rightarrow L_2}$) as relative coordinates (see figure 1). In order to ensure exact reproduction, we recalculated these relative coordinates alongside the proposed BAT coordinates and compared them to the data provided with [11]. Our comparison demonstrated an exact match. For further details, see the DeLinker publication [11].

2.4. Training

The model was trained under a VAE framework over 10 epochs, exclusively on the ZINC training set, using the Adam optimizer with a learning rate of 10^{-3} and a minibatch size of 16. The encoding of the nodes hidden states as well as the latent vector z encoding the ground-truth molecule were regularized via KL loss to follow standard normal distributions. A two-fold cross-entropy reconstruction loss was applied, one part measuring the error in the prediction of the atom types, the second part judging the sequence of bond-formation steps in order to reconstruct the ground truth molecule. Further details can be found in [11, 21].

2.5. Evaluation

For each of the fragment pairs in the triplet cuts of the test sets, i.e. 400 in the case of ZINC and 309 for CASF, 250 linkers were generated. This amounts to 100 000 and 77 250 linked molecules generated, respectively. We applied the standard metrics for molecular generation tasks, i.e. fractional validity, uniqueness and novelty. Validity was assessed by RDKit [39] being able to parse the generated SMILES [45] strings as connected molecules. Uniqueness was calculated by the cardinality of the (unique) set of generated molecules divided by the total number generated. Novelty describes the fraction of generated molecules which were not contained in the training set.

Furthermore, we assessed the generated molecules via the 2D filtering metrics given in subsection 2.3. SA [41] is a measure for the difficulty of physically synthesizing the molecule in the laboratory. Ring aromaticity amounts to the properties of a molecule constituting a drug. PAINS [42] assesses the reactivity of a compound by performing a knowledge-based analysis on its substructures.

Arguably the most significant estimator of the impact of our coordinate system is the models capability to recover the ground-truth linker from the original triplet cut, as the structural information provided to the generation process is derived from the respective ground-truth linker. In this manner, the model is conditioned to reproduce the ground-truth linker more frequently.

2.6. Information-theoretical analysis

Input features, referring to the coordinates in the present case, should be uncorrelated, i.e. decoupled, in order to assure efficient learning for the model. Therefore, the mutual information between the input features should be low [46]. In order to investigate the decoupling, we calculate pairwise mutual information values, given as [12–14, 17–19]

$$I(X,Y) = S(X) + S(Y) - S(X,Y) \tag{4}$$

with

$$\begin{aligned} S(X) &= -\sum_i p_i^x \ln\left(\frac{p_i^x}{J_i^x \Delta x}\right) \\ S(Y) &= -\sum_j p_j^y \ln\left(\frac{p_j^y}{J_j^y \Delta y}\right) \\ S(X,Y) &= -\sum_{i,j} p_{i,j}^{x,y} \ln\left(\frac{p_{i,j}^{x,y}}{J_i^x J_j^y \Delta x \Delta y}\right). \end{aligned} \tag{5}$$

Equation (5) refers to discretized differential entropy values. This means the underlying continuous probability densities are approximated by sampling to discrete histogram bins. The indices x and y designate an arbitrary coordinate, i.e. a bond, an angle, or a dihedral in the present case. p_i^x denotes the probability for the coordinate x to assume a value in bin i. If N samples for coordinate x are taken, and N_i of them fall into bin i, then $p_i^x = N_i/N$ (and analogous for p_j^y). Furthermore, $p_{i,j}^{x,y} = N_{i,j}/N$ if $N_{i,j}$ of the N samples taken fall into bin i for the x-coordinate and bin j for the y-coordinate in the according 2D histogram. Δx and Δy are the widths of the equally spaced bins. J_i^x is the Jacobian of coordinate x in the middle of bin i. Using the definition of the marginals

$$\begin{aligned} p_i^x &= \sum_j p_{i,j}^{x,y} \\ p_j^y &= \sum_i p_{i,j}^{x,y}, \end{aligned} \tag{6}$$

we can write equation (4) as

$$\begin{aligned} I(X,Y) &= \sum_{i,j} p_{i,j}^{x,y} \ln\left(\frac{p_{i,j}^{x,y}/(J_i^x J_j^y \Delta x \Delta y)}{p_i^x/(J_i^x \Delta x) * p_j^y/(J_j^y \Delta y)}\right) \\ &= \sum_{i,j} p_{i,j}^{x,y} \ln\left(\frac{p_{i,j}^{x,y}}{p_i^x p_j^y}\right). \end{aligned} \tag{7}$$

The last equality exhibits interesting features. First, the Jacobian determinants have canceled out, meaning the type of degree of freedom has become irrelevant. Furthermore, the bin sizes have vanished. This means that the mutual information calculated from discretized differential entropy values resembles discrete information in Shannon's [47] sense. Note, however, that the mutual information still is dependent on bin sizes; the probabilities $p_{i,j}^{x,y}$, p_i^x and p_j^y are affected by this choice. For example, let us denote the limit where the binning becomes binary, meaning the bin sizes are chosen so small that there is either no or only one data point in each bin. Assuming that both the x- as well as the y-marginal take on such a binary form, we can write for N data points:

$$\begin{aligned} I(X,Y) &= \sum_{i,j} p_{i,j}^{x,y} \ln\left(\frac{p_{i,j}^{x,y}}{p_i^x p_j^y}\right) \\ &= \sum_{p_{i,j}^{x,y}>0} \frac{1}{N} \ln\left(\frac{1/N}{(1/N)*(1/N)}\right) \\ &= \frac{N}{N} \ln(N) \\ &= \ln(N). \end{aligned} \tag{8}$$

Furthermore, note that we write the equations without the Boltzmann or Gas constant. Spatial entropies, as in the equation (5), do not bear physically meaningful units. The reason is the fact that the semi-classical entropy integral cannot be split in a manner that either the spatial or the momentum entropy bear physically meaningful units [14]. Upon taking entropy differences, however, the problematic terms cancel out [14]. This means that the mutual information in equations (4), (7) and (8) can indeed be multiplied with the Boltzmann or Gas constant to yield physical entropy values. Stated without such a physical constant, the units obtained here are natural units of information (nats, similar to bits, however, using the natural logarithm).

Table 1. Graph metrics. 2D quality criteria are compared for molecules generated using BAT coordinates versus the values from the DeLinker [11] publication. Values which are not given in the DeLinker [11] publication are marked via an 'N/A' entry. For the ZINC dataset, the DeLinker [11] ablation study values are included. The two 5+ test sets at the bottom constitute subsets of target linkers with at least five atoms in length. Best in category (row) values have been printed in bold font.

		No info	Distance only	DeLinker	BAT
Recovered	ZINC	74.5	78.3	79.0	**88.3**
	CASF	N/A	N/A	53.7	**56.3**
	ZINC 5+ atoms	N/A	N/A	67.0	**80.8**
	CASF 5+ atoms	N/A	N/A	29.8	**34.0**
Novel	ZINC	36.2	37.6	39.5	**39.6**
	CASF	N/A	N/A	51.0	**53.3**
	ZINC 5+ atoms	N/A	N/A	**49.4**	47.9
	CASF 5+ atoms	N/A	N/A	68.7	**69.1**
Valid	ZINC	97.0	**98.6**	98.4	98.2
	CASF	N/A	N/A	**95.5**	94.5
	ZINC 5+ atoms	N/A	N/A	98.1	**98.3**
	CASF 5+ atoms	N/A	N/A	**94.7**	93.1
Unique	ZINC	**51.2**	47.3	44.2	37.6
	CASF	N/A	N/A	**51.9**	47.1
	ZINC 5+ atoms	N/A	N/A	**61.0**	52.7
	CASF 5+ atoms	N/A	N/A	**72.9**	66.3
Pass all 2D filters	ZINC	89.9	90.2	89.8	**90.5**
	CASF	N/A	N/A	**81.4**	80.4
	ZINC 5+ atoms	N/A	N/A	84.1	**85.5**
	CASF 5+ atoms	N/A	N/A	**71.7**	70.3
Pass ring filter	ZINC	95.2	94.5	94.8	**95.5**
	CASF	N/A	N/A	N/A	92.5
	ZINC 5+ atoms	N/A	N/A	N/A	92.2
	CASF 5+ atoms	N/A	N/A	N/A	87.2
Pass SA filter	ZINC	95.1	**95.5**	95.3	**95.5**
	CASF	N/A	N/A	N/A	85.1
	ZINC 5+ atoms	N/A	N/A	N/A	93.4
	CASF 5+ atoms	N/A	N/A	N/A	77.6
Pass PAINS filter	ZINC	97.8	**98.4**	97.9	98.2
	CASF	N/A	N/A	N/A	98.0
	ZINC 5+ atoms	N/A	N/A	N/A	97.7
	CASF 5+ atoms	N/A	N/A	N/A	97.7

3. Results

When comparing different generative models for linker generation, arguably the most indicative metric for the quality of the provided coordinates is the rate of recovery of the ground-truth linker, as it quantifies the models capability to follow the supplied information. Table 1 shows that this metric improves across all test sets when using the full BAT coordinate system, with the most drastic enhancement for ZINC. The coordinate system carries information about the ground-truth linker. With higher quality information supplied, the model is expected to recover the original linker more frequently. This demonstrates that the model takes advantage of the augmented information by following the provided target geometry. Since the generation process is more directed, the uniqueness metrics drops, which serves as another indicator for the quality of the information content in our proposed coordinate system.

Considering the ablation study of DeLinker [11] on the ZINC test set (see table 1), the bulk of the geometric information obviously is contained in the distance coordinate. Compared to providing no relative coordinates at all, the distance takes the recovery rate from 74.5 to 78.3 percent. Adding the angle coordinate, i.e. considering the complete DeLinker [11] coordinate set, yields an additional gain of comparatively low 0.7 percent. Figure 3 provides insight on this outcome. When using a different random seed for RDKit to generate conformers, the DeLinker [11] distances are preserved rather robustly with a Pearson correlation of 0.86. The angles, on the other hand, show significant deviations. Their Pearson correlation to the angles of the conformers generated with the previous random seed is given as a relatively low value of 0.47. This

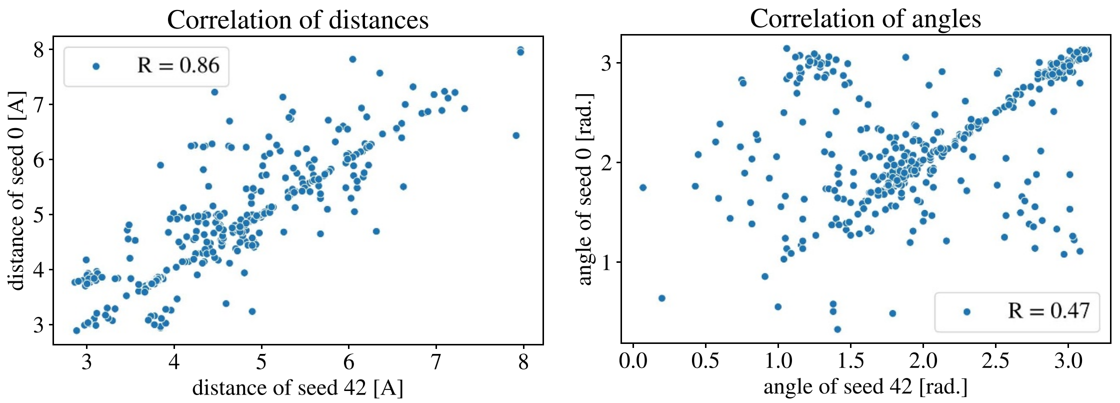

Figure 3. Robustness of DeLinker coordinates. The plots compare two random seeds used for generating conformers for the test set. The distance as well as the angle of the DeLinker coordinate set are plotted. The distances demonstrate a somewhat adequate robustness between random seeds, with Pearson correlation of 0.86 between seeds. For the angles, we observe a considerable fraction of outliers, decreasing the correlation to 0.47 only. The robustness to altering random seeds is proposed as a proxy for the quality of the provided structural information.

Table 2. BAT ablation study. The same metrics as in table 1 are investigated and the impact of the different types of coordinates is dissected. As was the case for the DeLinker [11] coordinates, the 2D filter and the valid criteria are generally high. Referring to the recovered metric, similar to table 1, the bulk of our information is carried in the distance coordinate, albeit the improvement appears significantly more distinct. Furthermore, the unique metric continues to behave opposite to the recovery metric, supporting the hypothesis of the coordinate information providing valuable guidance for the model.

		No info	Distance only	Distance and angles	BAT
ZINC	Recovered	74.5	84.5	87.0	**88.3**
	Novel	36.2	**40.6**	38.6	39.6
	Valid	97.0	97.5	**98.2**	**98.2**
	Unique	**51.2**	44.8	37.8	37.6
	Pass all 2D filters	89.9	89.8	**90.5**	**90.5**
	Pass ring filter	95.2	94.5	**95.6**	95.5
	Pass SA filter	95.1	95.1	94.9	**95.5**
	Pass PAINS filter	97.8	98.0	**98.2**	**98.2**

indicates that the reliability of the angular information is low compared to the distance information, which attributes to only minor gains when feeding them to the model.

In order to further investigate the information content in the relative coordinates proposed here, analogue ablation studies were performed for the BAT coordinates. Table 2 lists the results.

Similar to the DeLinker [11] coordinate system, the bulk of the BAT information is carried in the distance coordinate. However, the BAT distance ($|\vec{r}_{L_1 \to L_2}|$ in figure 1) takes the recovered metric from 74.5 to 84.5 (table 2), as opposed to only 78.3 (table 1) for the DeLinker [11] distance $|\vec{r}_{E_1 \to E_2}|$. Given the similarity of the two distances, this discrepancy appears rather drastic at a first glance. The reason arguably lies in the fact that the DeLinker [11] distance fails to decouple from the angular and dihedral coordinate systems. For the 6 BAT coordinates in figure 1, one can vary each of them while keeping the others constant. However, any such variation will change the DeLinker [11] distance $|\vec{r}_{E_1 \to E_2}|$. Furthermore, varying the BAT angles or the BAT dihedral will change the DeLinker [11] angle. When comparing the distances in figure 3 to those in figure 4, the effect of this coupling becomes evident. The DeLinker [11] distance shows a Pearson correlation of 0.86 between coordinates of conformers generated with different random seeds. The decoupled BAT distance, on the other hand, demonstrates excellent robustness to the choice of random seed with a Pearson correlation of 0.99. This explains the arguably drastic improvement of the recovered metric by 10 percent using the BAT distance only.

Feeding the model the BAT angles additionally, we gain another 2.5 percent, as opposed to 0.7 for the DeLinker [11] angle. Comparing the correlation plots, BAT provides two angles with a Pearson correlation of 0.67 and 0.72 versus one angle with 0.47 for DeLinker [11]; the BAT angles are decoupled from the distance and dihedral system.

Figure 4 shows that the dihedral angle information is of low quality. Interestingly, even given this low quality information, the model improves the recovered metric by 1.3 percent. Note that this value exceeds the angular improvement of the DeLinker [11] coordinates (0.7) almost by a factor of two. The reason, as above, arguably lies in the decoupling from the distance and angular system; while the information is certainly

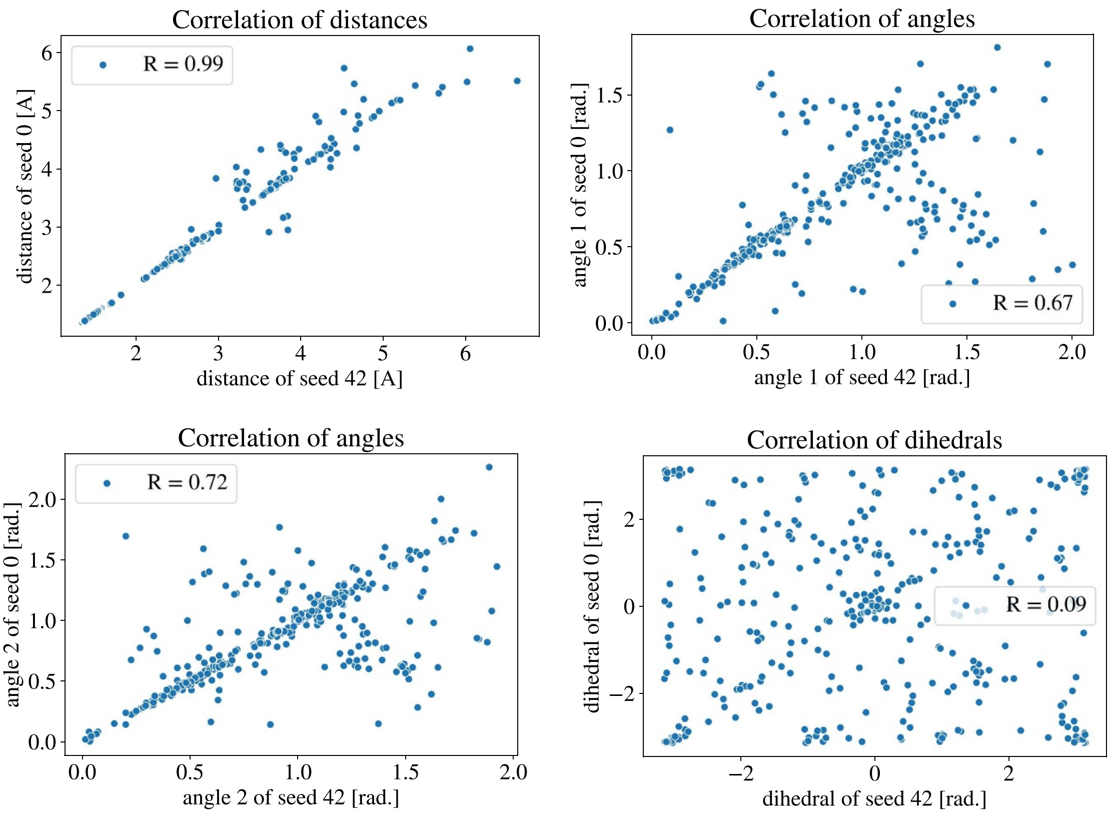

Figure 4. BAT coordinate quality. Analogous plot to figure 3 for the proposed BAT coordinates. The robustness to varying random seeds for the distances appears excellent with a Pearson $R = 0.99$. Compared to figure 3, both our angles demonstrate enhanced reliability. Referring to the dihedrals, the estimated information content is considerably low.

non-reliable on its own, it is indeed orthogonal, i.e. non-redundant, with respect to the other coordinates. Altogether, the BAT coordinate system takes the recovery metric to 88.3 percent, which is an improvement of 9.3 percent over the DeLinker [11] coordinate system.

Finally, we performed an information-theoretical analysis comparing the two input coordinate sets. Given that the DeLinker [11] coordinates are given as a distance and angle pair, we compare the mutual information using the training set between the Delinker [11] coordinates with the mutual information between the BAT distance and either BAT angle (figure 5). We find that, with few exceptions at large bin sizes where values have not converged, the BAT coordinates exhibit lower mutual information, which demonstrates an increased amount of decoupling. Given these insights on an information-theoretic level, we suggest the presented graphs as strong evidence for the decoupling of the coordinate system as a major factor for the improvements.

4. Discussion and conclusion

Inspired by recent pioneering work [11], we have demonstrated a considerable beneficial effect of the application of a well-behaved coordinate system on machine learning-based molecular linker generation. The enhancements of our proposed relative coordinate system, which roots in the BAT coordinate formalism, were established by demonstrating the improvement of such a coordinate system in the framework of [11]: our approach allowed to decrease the number of test set examples for which the ground-truth linker could not be recovered by roughly one half (a reduction of 44.3 percent). We performed comparative analyses on various indicative aspects: first, using common metrics for generative models as well as molecular graph evaluation, and second, ablation studies on the included coordinate types which allowed identifying the coordinates which provide the most valuable information to the model. We furthermore investigated the reasons for the performance of the different coordinates by performing a robustness analysis with respect to the conformer generation process given different random seeds. By means of information-theoretical calculations, we compared the amount of decoupling in the two input coordinate sets. The results support the hypothesis that the decoupled nature of our presented coordinate system plays a major role for the improved performance of the generative process.

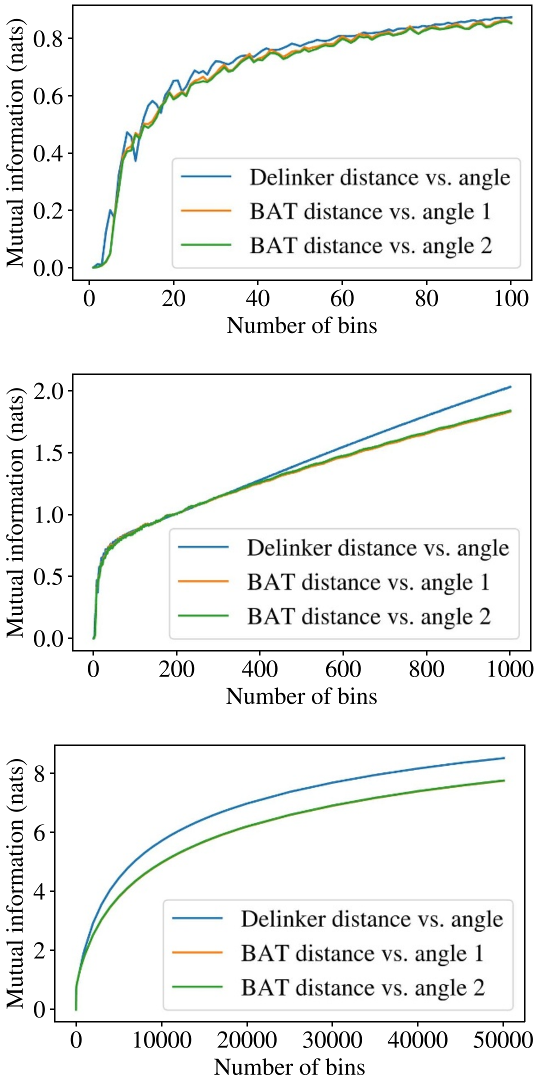

Figure 5. Mutual information comparison. The plots compare the mutual information between the DeLinker [11] distance and angle as well as between the BAT distance and bond angles on the training set of ground-truth molecules. Since the mutual information is dependent on the bin size (see section 2), the plots show the mutual information as a function of the granularity in terms of the number of bins used. Different regimes of the number of bins are shown. Note that the mutual information curves for both our angles lie on top of each other. With the exception of rare events in the regime of low number of bins (top graph), our coordinates exhibit lower mutual information, i.e. enhanced decoupling.

Our findings highlight the advantage of utilizing structural information in future models for molecular fragment linking. Given the considerable improvements demonstrated, we propose the presented coordinate system as a standard technique for linking molecular fragments.

The application of structural information in models of linker generation is of high relevance for fragment-based drug discovery. While some progress has been made in understanding the role of the linker in compounds designed for targeted protein degradation [8], many aspects remain poorly understood. Supplying enhanced structural information to linker generation methods can lead to better *in silico* proposals here, allowing to focus *in vitro* evaluation on more promising candidates. The relative coordinate system presented in this work constitutes a first step towards more 3D aware models, which may take into account not only the relative positions of the fragments, but also (e.g.) a desired shape of the linker. As the properties of the linker play an important role in defining the overall compound efficacy, a better understanding of its geometry constitutes a key ingredient for designing improved pharmaceuticals.

Data availability statement

No new data were created or analysed in this study.

Acknowledgments

We thank Anton A Polyansky for a critical reading of the manuscript. All authors are affiliated with Celeris Therapeutics GmbH. Christopher Trummer and Noah Weber are C-level officials in Celeris Therapeutics GmbH and are holding equities. Celeris Therapeutics GmbH has filed a patent application with the Number PCT/EP2021/025372, which uses the presented enhanced coordinate system as part of its pipeline.

Appendix. Transformation from BAT to anchored Cartesian coordinates

In order to demonstrate the completeness of the proposed BAT coordinate system (equation (3)), the back-transformation to anchored Cartesian coordinates [26, 27] (equation (1)) is given as follows.

$$
\begin{aligned}
L_1^x &= |\vec{b_1}| \\
L_2^x &= L_1^x + |\vec{b_2}| \cos(\alpha_1) \\
L_2^y &= |\vec{b_2}| \sin(\alpha_1) \\
\vec{r}_{E_2} &= \vec{r}_{L_2} + \vec{b_3}' \cos(\phi) + \left(\frac{\vec{b_2}}{|\vec{b_2}|} \times \vec{b_3}' \right) \sin(\phi) \\
&\quad + \left(\frac{\vec{b_2}^{\mathsf{T}}}{|\vec{b_2}|} \cdot \vec{b_3}' \right) [1 - \cos(\phi)] \frac{\vec{b_2}}{|\vec{b_2}|}
\end{aligned}
$$

with

$$
\vec{b_3}'^{\mathsf{T}} = \Big(|\vec{b_3}| \cos(\alpha_1 + \alpha_2),\ \ |\vec{b_3}| \cos(\alpha_1 + \alpha_2),\ \ 0 \Big). \tag{9}
$$

Here, $\vec{b_3}'$ represents $\vec{b_3}$ for the case $\phi = 0$. In order to obtain $\vec{b_3}$ from $\vec{b_3}'$, we rotate $\vec{b_3}'$ around an axis parallel to $\vec{b_2}$ by an angle ϕ. Following previous work [31], this coordinate transformation is applied by using Rodrigues' rotation formula. Note that equation (9) is never explicitly calculated. It is given here for the sole purpose of demonstrating the equivalence of the BAT and internal Cartesian coordinate systems.

ORCID iDs

Markus Fleck https://orcid.org/0000-0002-8648-2164
Michael Müller https://orcid.org/0000-0001-7025-5090
Christopher Trummer https://orcid.org/0000-0002-4985-9962

References

[1] Polishchuk P G, Madzhidov T I and Varnek A 2013 Estimation of the size of drug-like chemical space based on GDB-17 data *J. Comput. Aided Mol. Des.* **27** 675–9

[2] Reymond J-L, van Deursen R, Blum L C and Ruddigkeit L 2010 Chemical space as a source for new drugs *MedChemComm* **1** 30

[3] Xuanyi Li, Yinqiu X, Yao H and Lin K 2020 Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors *J. Cheminf.* **12** 1–13

[4] Mullard A 2017 The drug-maker's guide to the galaxy *Nature* **549** 445–7

[5] Dale B, Cheng M, Park K-S, Kaniskan H Ü, Xiong Y and Jin J 2021 Advancing targeted protein degradation for cancer therapy *Nat. Rev. Cancer* **21** 638–54

[6] Troup R I, Fallan C and Baud M G J 2020 Current strategies for the design of PROTAC linkers: a critical review *Explor. Target. Antitumor Ther.* **1** 273–312

[7] Cecchini C, Pannilunghi S, Tardy S and Scapozza L 2021 From conception to development: investigating PROTACs features for improved cell permeability and successful protein degradation *Front. Chem.* **9** 672267

[8] Bemis T A, La Clair J J and Burkart M D 2021 Unraveling the role of linker design in proteolysis targeting chimeras *J. Med. Chem.* **64** 8042–52

[9] Ichihara O, Barker J, Law R J and Whittaker M 2011 Compound design by fragment-linking *Mol. Inf.* **30** 298–306

[10] Bienstock R J 2015 *Computational Methods for Fragment-Based Ligand Design: Growing and Linking* (New York: Springer) pp 119–35

[11] Imrie F, Bradley A R, van der Schaar M and Deane C M 2020 Deep generative models for 3D linker design *J. Chem. Inf. Model.* **60** 1983–95

[12] Killian B J, Yundenfreund Kravitz J and Gilson M K 2007 Extraction of configurational entropy from molecular simulations via an expansion approximation *J. Chem. Phys.* **127** 024107

[13] Killian B J, Kravitz J Y, Somani S, Dasgupta P, Pang Y-P and Gilson M K 2009 Configurational entropy in protein–peptide binding: computational study of Tsg101 ubiquitin E2 variant domain with an HIV-derived PTAP nonapeptide *J. Mol. Biol.* **389** 315–35

[14] Hnizdo V and Gilson M K 2010 Thermodynamic and differential entropy under a change of variables *Entropy* **12** 578–90

[15] Baron R, van Gunsteren W F and Hünenberger P H 2006 Estimating the configurational entropy from molecular dynamics simulations: anharmonicity and correlation corrections to the quasi-harmonic approximation *Trends Phys. Chem.* **11** 87–122

[16] King B M, Silver N W and Tidor B 2012 Efficient calculation of molecular configurational entropies using an information theoretic approximation *J. Phys. Chem.* B **116** 2891–904

[17] Fleck M, Polyansky A A and Zagrovic B 2016 PARENT: a parallel software suite for the calculation of configurational entropy in biomolecular systems *J. Chem. Theory Comput.* **12** 2055–65

[18] Fleck M and Zagrovic B 2019 Configurational entropy components and their contribution to biomolecular complex formation *J. Chem. Theory Comput.* **15** 3844–53

[19] Numata J and Knapp E-W 2012 Balanced and bias-corrected computation of conformational entropy differences for molecular trajectories *J. Chem. Theory Comput.* **8** 1235–45

[20] Fleck M, Wieder M and Boresch S 2021 Dummy atoms in alchemical free energy calculations *J. Chem. Theory Comput.* **17** 4403–19

[21] Liu Q, Allamanis M, Brockschmidt M, and Gaunt A 2018 Constrained graph variational autoencoders for molecule design Bengio S *Advances in Neural Information Processing Systems* vol 31, ed H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi and R Garnett (Curran Associates, Inc.)

[22] Jin W, Yang K, Barzilay R and Jaakkola T 2019 Learning multimodal graph-to-graph translation for molecule optimization *Int. Conf. on Learning Representations*

[23] Yujia Li, Tarlow D, Brockschmidt M, and Zemel R S 2016 Gated graph sequence neural networks *Int. Conf. on Learning Representations* ed Y Bengio and Y LeCun

[24] Kingma D P and Welling M 2014 Auto-encoding variational bayes *Int. Conf. on Learning Representations* ed Y Bengio and Y LeCun

[25] Zhu J-Y, Zhang R, Pathak D, Darrell T, Efros A A, Wang O, and Shechtman E 2017 Toward multimodal image-to-image translation *Advances in Neural Information Processing Systems* ed I Guyon, U von Luxburg, S Bengio, H M Wallach, R Fergus, S V N Vishwanathan and R Garnett pp 465–76

[26] Potter M J and Gilson M K 2002 Coordinate systems and the calculation of molecular properties *J. Phys. Chem.* A **106** 563–6

[27] Chang C-E, Potter M J and Gilson M K 2003 Calculation of molecular configuration integrals *J. Phys. Chem.* B **107** 1048–55

[28] Herschbach D R, Johnston H S and Rapp D 1959 Molecular partition functions in terms of local properties *J. Chem. Phys.* **31** 1652–61

[29] Pitzer K S 1946 Energy levels and thermodynamic functions for molecules with internal rotation: II. Unsymmetrical tops attached to a rigid frame *J. Chem. Phys.* **14** 239–43

[30] Gā N and Scheraga H A 1976 On the use of classical statistical mechanics in the treatment of polymer chain conformation *Macromolecules* **9** 535–42

[31] Parsons J, Holmes J B, Rojas J M, Tsai J and Strauss C E M 2005 Practical conversion from torsion space to Cartesian space for in silico protein synthesis *J. Comput. Chem.* **26** 1063–8

[32] Gordon M S and Pople J A 1968 Approximate self-consistent molecular-orbital theory. VI. INDO calculated equilibrium geometries *J. Chem. Phys.* **49** 4643–50

[33] Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, Sattar A, Yang Y and Zhou Y 2015 Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning *Sci. Rep.* **5** 11476

[34] Haiou Li, Hou J, Adhikari B, Lyu Q and Cheng J 2017 Deep learning methods for protein torsion angle prediction *BMC Bioinform.* **18** 417-1–13

[35] Gao J, Yang Y and Zhou Y 2018 Grid-based prediction of torsion angle probabilities of protein backbone and its application to discrimination of protein intrinsic disorder regions and selection of model structures *BMC Bioinform.* **19** 29-1–8

[36] Gómez-Bombarelli R *et al* 2018 Automatic chemical design using a data-driven continuous representation of molecules *ACS Cent. Sci.* **4** 268–76

[37] Sterling T and Irwin J J 2015 ZINC 15—ligand discovery for everyone *J. Chem. Inf. Model.* **55** 2324–37

[38] Minyi S, Yang Q, Du Y, Feng G, Liu Z, Li Y and Wang R 2019 Comparative assessment of scoring functions: the CASF-2016 update *J. Chem. Inf. Model.* **59** 895–913

[39] Landrum G 2018 RDKit: open-source cheminformatics (available at: www.rdkit.org/)

[40] Hussain J and Rea C 2010 Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets *J. Chem. Inf. Model.* **50** 339–48

[41] Ertl P and Schuffenhauer A 2009 Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions *J. Cheminform.* **1** 8

[42] Baell J B and Holloway G A 2010 New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays *J. Med. Chem.* **53** 2719–40

[43] Halgren T A 1996 Merck molecular force field. I. Basis, form, scope, parameterization and performance of MMFF94 *J. Comput. Chem.* **17** 490–519

[44] Halgren T A 1999 MMFFf VI. MMFFf94s option for energy minimization studies *J. Comput. Chem.* **20** 720–9

[45] Weininger D 1988 Smiles, a chemical language and information system. 1. Introduction to methodology and encoding rules *J. Chem. Inf. Comput. Sci.* **28** 31–36

[46] Brown G 2009 A new perspective for information theoretic feature selection *Proc. 12th Int. Conf. on Artificial Intelligence and Statistics* ed D van Dyk and M Welling (PMLR) pp 49–56

[47] Shannon C E 1948 A mathematical theory of communication *Bell Syst. Tech. J.* **27** 379–423