COATI: MULTI-MODAL CONTRASTIVE PRE-TRAINING FOR REPRESENTING AND TRAVERSING CHEMICAL SPACE

A PREPRINT

Benjamin Kaufman **Edward C. Williams** **Carl Underkoffler** **Ryan Pederson**

Narbe Mardirossian **Ian Watson** **John Parkhill**
Terray Therapeutics, Inc

August 25, 2023

ABSTRACT

Creating a successful small molecule drug is a challenging multi-parameter optimization problem in an effectively infinite space of possible molecules. Generative models have emerged as powerful tools for traversing data manifolds comprised of images, sounds, and text, and offer an opportunity to dramatically improve the drug discovery and design process. To create generative optimization methods that are more useful than brute-force molecular generation and filtering via virtual screening, we propose that four integrated features are necessary: large, quantitative datasets of molecular structure and activity, an invertible vector representation of realistic accessible molecules, smooth and differentiable regressors that quantify uncertainty, and algorithms to simultaneously optimize properties of interest. Over the course of 12 months, Terray has collected a dataset of 2 billion quantitative binding measurements, which directly motivates multi-parameter generative optimization of molecules conditioned on this data. To this end, we present COATI[1], a pre-trained, multi-modal encoder-decoder model of druglike chemical space. COATI is constructed without any human biasing of features, using contrastive learning from text and 3D representations of molecules to allow downstream use with structural models. We demonstrate that COATI possesses many of the desired properties of a universal molecular embedding: fixed-dimension, invertibility, autoencoding, accurate regression, and low computation cost. Finally, we present a novel metadynamics algorithm for generative optimization using a small subset of our proprietary data collected for a model protein, Carbonic Anhydrase, designing molecules that satisfy the multi-parameter optimization task of potency, solubility, and druglikeness. This work sets the stage for fully-integrated generative molecular design and optimization for small molecules.

Keywords contrastive learning, generative optimization, molecular generation, small molecules, drug design, drug discovery

1 Introduction

The space of druglike small molecules is estimated to contain over 10^{60} unique structures (Reymond, 2015). Because of its sheer size, brute force search over the entire space is impossible. Typical drug discovery campaigns rely on a two-step approach for virtually evaluating select parts of this space for both hit discovery efforts and molecular optimization efforts: molecular generation and property prediction/filtering. A variety of approaches to generation are commonly used: reaction-based or combinatorial enumeration, evolutionary algorithms, molecular expertise/intuition, scaffold replacement, and others. There are also "pre-generated" virtual chemical spaces such as Enamine REAL (Grygorenko et al., 2020) and WuXi Galaxi (Xu, 2021) that can be explored. However, for hit-to-lead and lead optimization efforts, de-novo generation is commonly used to explore focused chemical spaces that are not found in commercially-available collections. Once chemical spaces of interest are produced, a variety of property prediction models can be used to triage costly experimental measurements. These methods can be split into two

[1]Contrastive Optimization for Accelerated Therapeutic Inference

categories: machine learning (ML)-based methods and physics-based methods. Ligand-based ML methods for small molecule property prediction have three basic requirements: experimental molecular training data (e.g., IC50s from a biochemical activity assay), a method for featurizing or representing small molecules (e.g., molecular fingerprints), and a supervised learning algorithm for classification or regression (e.g., an XGBoost regressor (Chen and Guestrin, 2016)). On the other hand, physics-based approaches typically require little to no experimental input and rely on quantum mechanics (i.e., Density Functional Theory), classical mechanics (i.e., force fields), or other related methods to calculate properties such as conformational strain, solubility, or target-ligand binding (Wang et al., 2015; Bannwarth et al., 2019).

While this two-step approach to hit discovery and molecular optimization has been successfully used in many drug discovery programs, a one-step generative method that can simultaneously explore chemical space while optimizing or constraining properties of interest would be far more computationally efficient for hit-to-lead and lead optimization. This approach of generative molecular optimization requires four fundamental components: a) an iterative source of quantitative, experimental data at scale, b) a molecular featurization or representation that is decodable and predictive of molecular properties – a primary focus of this work, c) smooth and differentiable regressors with built-in uncertainty quantification, and d) chemical space exploration methods that can be used to directly explore and optimize on the surface of a regressor. Due to the appeal of this approach, many generative models for molecules have been reported in both text (Reidenbach et al., 2023; Seidl et al., 2023; Blaschke et al., 2020; Winter et al., 2019), graph (Bengio et al., 2021; Liu et al., 2023; Vignac et al., 2022) and 3D (Pinheiro et al., 2023) modalities.

A method for molecular generation that has been gaining popularity in recent years as an alternative to the conventional methods mentioned earlier is the use of unsupervised learning algorithms to generate a decodable latent space vector representation of molecules that can be explored directly in the vector representation and subsequently decoded to a valid molecule. These pre-trained, generalizable encoders have become a popular molecular design tool in recent years (Yang et al., 2021; Kim et al., 2022; Masters et al., 2022). However, these models may operate on different chemical representations with no clear optimal choice. Contrastive learning approaches are able to integrate several data modalities, can boost robustness on downstream tasks, and have been shown to be successful in multiple fields (Radford et al., 2021a; Stärk et al., 2021; Xue et al., 2022). We explore a scheme that uses contrastive learning of multiple molecular modalities, and our experiments show that this strategy leads to broadly applicable and robust representations. More generally, we seek a generative foundation model of small molecules that decouples conditional generation from fine-tuning of the foundation model and provides a path forward for future multi-modal representation learning advances. To this effect, we present COATI, a novel and practical method for generative molecular design that can be used with *any* set of molecular properties that can be expressed by a differentiable model. We rely on a novel multi-modal encoder-decoder scheme for structures that is a competitive encoder for a variety of molecule regression tasks. We achieve this by aligning 2D (text) and 3D (point) representations of molecules, and simultaneously training a generative transformer decoder which recovers a molecule from either input. We demonstrate that contrastive pre-training leads to excellent regression performance vs. fingerprints, encoder-only models, and other decodable representations. We report architecture variations of the model, and quantify likelihoods of decoding and generating various molecular spaces. In direct analogy with recent developments in text-to-image generative models (Radford et al., 2021a; Ramesh et al., 2021; Rombach et al., 2021; Saharia et al., 2022) that inspired this work, a common latent space for molecular representations which is space-agnostic (Wellawatte et al., 2022) has many practical uses.

We envision that this generative design method will be used in concert with novel, high-throughput experimental methods capable of iteratively generating data relevant to molecular discovery (e.g., target-ligand binding measurements). We anticipate that our decodable molecular representation, COATI, along with our proposed metadynamics-inspired algorithm for molecular design and optimization will provide a useful template for future developments in practical generative design. This paper initially focuses on the development and assessment of COATI, and concludes with a real-world application of generative molecular optimization, using a small subset of the Terray platform data for a model protein (Carbonic Anhydrase) to generate molecules with optimized potency while satisfying multiple embedded property constraints.

2 Prior Works

2.1 Molecular Representations

Small molecules can be represented by a variety of methods including strings, binary fingerprints, property-based descriptors, and 3D coordinates. The most common string-based representation for molecules is called Simplified Molecular-Input Line-Entry System (SMILES) (day). We will also consider SELFIES (Krenn et al., 2020, 2022), which are by construction always valid. Both representations encode molecular graph topology into a string of text, and can be conveniently used with machine learning methods designed to process text for the purposes of property prediction (Honda et al., 2019) or autoregressive generation (Olivecrona et al., 2017).

Traditional methods of vectorizing molecular structures for input into machine learning models have used graph topology hashing (Morgan, 1965; Rogers and Hahn, 2010), substructure queries (Durant et al., 2002), or pairwise iteration of atom distances (Carhart et al., 1985; Capecchi et al., 2020). These "molecular fingerprints" have been used for

decades to perform Quantitative Structure-Activity Relationship (QSAR) modeling (Muratov et al., 2020), virtual screening, and similarity search across large chemical databases (Maggiora et al., 2014; Muegge and Mukherjee, 2016). Traditional approaches are still widely used, and often have significant practical advantages for small datasets.

Recent research has sought to replace engineered features with representations learned directly from data. End-to-end learning methods have taken the form of supervised learning directly on molecular graphs using Graph Neural Networks (Gilmer et al., 2017), unsupervised learning on graphs or SMILES strings via autoencoders (Gómez-Bombarelli et al., 2018; Jin et al., 2018), GANs (De Cao and Kipf, 2018), or autoregressive pre-training (Honda et al., 2019). Learned molecular representations have shown promising results approximating quantum chemical calculations (Smith et al., 2017; Schütt et al., 2017; Schütt et al., 2017; Yao et al., 2017; Gilmer et al., 2017), predicting products of chemical reactions (Schwaller et al., 2021), and performing virtual screening after training on DNA-Encoded Library (DEL) data (McCloskey et al., 2020). Neural network architectures designed to be invariant or equivariant to transformations on Euclidean space (Thomas et al., 2018a; Fuchs et al., 2020; Satorras et al., 2021a) perform well simulating many-body systems and predicting properties of molecular configurations.

2.2 Generative Models for Molecules

Traditional Monte-Carlo algorithms and related approaches are able to sample plausible molecular structures either unconditionally, or with constraints (Gómez-Bombarelli et al., 2018; Jin et al., 2018). Data-driven methods have contributed new paths to molecular generation (Zhou et al., 2019), which can broadly be divided between autoregressive and one-shot approaches. Autoregressive generators build up a molecule step-wise, leveraging information from previous steps. The most common autoregressive models are text models (Olivecrona et al., 2017; Ahmad et al., 2022; Chilingaryan et al., 2022; Ross et al., 2022; Lee and Nam, 2022; Winter et al., 2019), but autoregressive graph models have also been explored (Shi et al., 2020). Winter et al. (2019) provides decodability as well as a secondary regression objective during pre-training, which we benchmark against our model architectures. Another approach is GFlowNets, which try to emulate autoregressive generation while technically being one-shot (Bengio et al., 2021). One-shot approaches that have been tried in this space include GANs (De Cao and Kipf, 2018), normalizing flows (Satorras et al., 2021b), and diffusion models (Vignac et al., 2022; Satorras et al., 2021b).

2.3 Unsupervised Contrastive Pre-training

Contrastive learning is a self-supervised learning paradigm that produces input space embeddings by training a model to match pairs of data points, either across different input modalities (Radford et al., 2021b) or augmentations of input data (Zbontar et al., 2021a), described later in Section 3.2.2. The CLIP (Contrastive Language-Image Pre-training) architecture uses a cross-entropy loss between encodings of different modalities and achieves state-of-the-art performance on many zero-shot learning tasks (Radford et al., 2021a). Other work has used graph contrastive learning (You et al., 2020) to learn molecular representations directly from 2D atomic connectivity graphs. There have also been several works which pre-train a mixed 2D/3D representation, although none to our knowledge based on an end-to-end contrastive loss with a decodable representation. Chen et al. used weighted graphs to incorporate 3D information (Chen et al., 2021). Liu et al. reported GraphMVP (Liu et al., 2021), which used an InfoNCE loss to align 2D and 3D graph representations (described in more detail in Section 3.2.1). Stärke et al. also rely on the InfoNCE loss to produce a joint 2D/3D encoder without decoding capability (Stärk et al., 2021). We experiment with this loss function as well as a related loss (Zbontar et al., 2021b), in addition to the added decoding ability that our model provides. Zhu et al. (2022) has reported pre-training with a mixed 2D/3D encoding and autoregressively generates 3D structures from their embedding. However, the point encoder used in that work is not equivariant and the learning objective invokes molecular substructures. Seidl et al. (2023) use contrastive learning to align embedding spaces of SMILES strings and natural language describing scientific assays, finding useful improvements in predictive performance on regression tasks. We focus instead on aligning multiple spaces of molecule structure data that are more intrinsic, but the addition of more modes into the joint representation space is a reasonable extension.

2.4 Related General Purpose Embeddings

Resting on the developments above, several methods occupying a similar niche to COATI will be discussed in the results: CLAMP (Seidl et al., 2023), MegaMolBART (mmb, 2022), ChemGPT (Frey et al., 2022), ChemBERTa MTR (Ahmad et al., 2022), and the model from Winter et al. (2019) which we will reference as CDDD. Table 1 summarizes the features of related model architectures discussed later in our linear probe regression results (Section 4.3)). Generative models can decode to molecules from their embeddings, which enables molecular design without filtering. Fixed dimensionality allows for simpler generative optimization on an embedding space using the encoder as a foundation model for regressors. We consider contrastive objectives desirable, because they can be scaled to large datasets without supervision and do not introduce bias towards property datasets. An alternative approach taken by models such as CDDD and ChemBERTa MTR is to add expressive power for features known to be important for druglike space to the learning objective. This gives good performance on available datasets, but calls into question whether features important to every downstream task have been chosen. Models with a contrastive loss or built-in bias for known drug features outperform encoder/decoders without these features. Finally, embedding from a 3D point representation of molecules enables combination of the encoder-decoder with 3D generative

Property	COATI	CLAMP	CDDD	MegaMolBART	ChemGPT	ChemBERTa MTR
Fixed Embed Dim.	✓	✓	✓			
Generative	✓		✓	✓	✓	
No Properties in Training	✓			✓	✓	
Contrastive Loss	✓	✓				
Text Input	✓	✓	✓	✓	✓	✓
3D Input	✓					

Table 1: COATI's features were chosen for unbiased generative molecular design; choices made by related methods are summarized here.

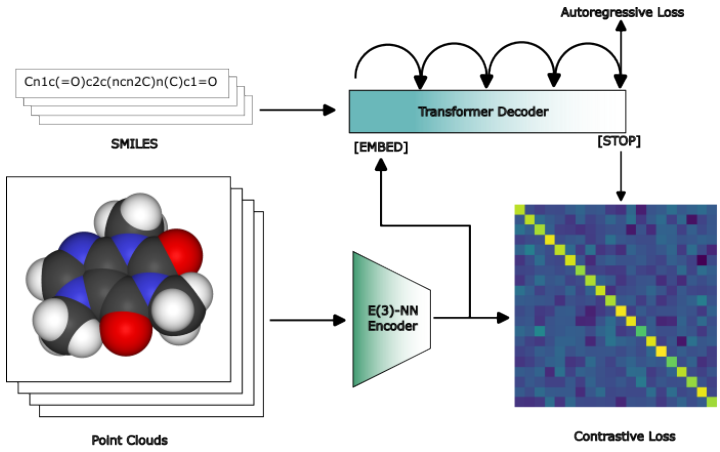

Figure 1: COATI is trained by jointly optimizing a contrastive loss with an autoregressive loss, producing a point cloud encoder and a SMILES transformer able to encode and decode.

models. For example, a generative diffusion model can be conditioned with such a model (Schneuing et al., 2023), although leveraging this feature is a topic for future work.

3 Methods

We seek a common representation for both SMILES and 3D molecular structures that can be used as input to predictive models, and is able to be decoded to generate new molecules.

3.1 Network Architecture

3.1.1 Point Encoder

For our 3D point encoder, we use the Welling group's E(3)-equivariant GNN (referred to as E(3)-GNN throughout this work) (Satorras et al., 2021a), a message passing network, because of its computational affordability relative to models with spherical tensor features (Unke et al., 2021; Thomas et al., 2018b; Batatia et al., 2022). Note that the use of an E(3)-equivariant point encoder implies that the version of the COATI latent space presented explicitly discards chirality information. This design choice can be relaxed by using an SE(3)-equivariant point encoder, or by tacking on a chiral spherical tensor invariant (Osipov et al., 1995) onto the point encoder. Atom nodes are embedded by two one-hot encoded vectors passed into a linear layer, for the row and period of the periodic table in which the element occurs (which we refer to as "periodic-one-hot" encoding). This node encoding improves loss over naïve embeddings per element, presumably because of an inductive bias where underrepresented elements can borrow from periodic relatives. Messages are differentiably masked beyond $r_c = 12$ by multiplying them by the cubic polynomial $\phi(r) = 1+(-3/2)r_c^{-2}r^2+(1/2)r_c^{-3}r^3$.

3.1.2 Text Encoder

COATI uses a rotary transformer (Su et al., 2021) to perform text encoding and decoding, and takes the representation of the `[STOP]` token as the embedding of SMILES or graph. We experiment with several schemes for tokenizing SMILES strings (Radford et al., 2019), all of which involve a basic trie encoder. We perform a token compression by recursively counting and appending the most common token pairs, starting from single characters. The process is repeated until the frequency of new composite tokens falls below a hand-tuned threshold, with some care taken to ensure the vocabulary retains irreducible tokens needed to span chemical space. We experiment with multiple tokenization schemes. We experimented with a vocabulary that only contained closed parenthetical groups and another that uses the SELFIES chemical representation (Krenn et al., 2020). SELFIES are constructed so that they

always translate to valid molecules. Additional details are described in Appendix 8.4. The mean token string length of a training example including all augmentations is only ∼15 tokens due to the significant compression afforded by pair tokens.

3.2 Learning Objective

We optimize a contrastive loss and autoregressive cross-entropy together in an end-to-end fashion (Fig. 1). Given a batch of SMILES/point cloud pairs of size K, each instance pair $(x_{\text{smiles}}, x_{\text{point}})$ has x_{smiles} augmented with an `[EMBED]` token with a 90% probability. We then select a representation vector to be injected into the `[EMBED]` token, inspired by the ClipCap injection procedure from Mokady et al. (2021).

50% of augmented token strings are selected for SMILES injection and 50% for point injection. Any SMILES injection pair gets a third representation x_{base}, the `[STOP]` token embedding from the SMILES transformer without augmentations applied to the input string. x_{point} is fed through the E(3)-GNN encoder head, then mapped through a linear layer to produce embedding z_p. If the instance was selected for point injection, the `[EMBED]` token embedding in x_{smiles} is replaced by z_p. If an instance is selected for SMILES injection, x_{base} is run through the transformer and the `[STOP]` token's representation is mapped through a linear layer to produce embedding z_{base} and the `[EMBED]` token embedding in x_{smiles} is replaced by z_{base}. Regardless of injection, x_{smiles} is then fed into the transformer with the usual softmax output and autoregressive entropy loss. The `[STOP]` token's representation is fed into a linear layer to produce z_s. The joint objectives are intended to allow the model to both *encode* and *decode* molecule structures, with the ability to decode from a point cloud to a SMILES string and to autoencode SMILES strings. We emphasize that the encoder and decoder are trained end-to-end with gradient flow through generation of the embedding token, and in our experiments this is necessary for training.

We experiment with two contrastive losses: the InfoNCE loss (Oord et al., 2018) and a cross-correlation loss we refer to as the "Barlow" loss (Zbontar et al., 2021b).

3.2.1 InfoNCE

For a given batch of size b with index i where $z_{s,i}$ is the d-dimensional embedding of a SMILES string, and $z_{p,i}$ is the d-dimensional embedding of a molecular point cloud, $\mathcal{L}_{\text{InfoNCE}}$ is defined as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2b}\sum_i \left(\ln \frac{\exp(z_{s,i}^\top z_{p,i})}{\sum_{j=0}^{K} \exp(z_{s,i}^\top z_{p,j})} + \ln \frac{\exp(z_{s,i}^\top z_{p,i})}{\sum_{j=0}^{K} \exp(z_{s,j}^\top z_{p,j})} \right) \quad (1)$$

In short, this maximizes the cosine similarities of matched embeddings from the two heads and minimizes the similarities of unmatched embeddings, by minimizing the entropy classifying each row to belong to the diagonal.

3.2.2 "Barlow" Cross-Correlation

We also experimented with a loss derived from Zbontar et al. (2021a), which we refer to as the "Barlow" loss. Rather than focusing on moving non-matched vectors "far apart" in terms of the cosine distance, this loss minimizes the cross-correlation between off-diagonal vector components of the pairs in a batch. We find empirically that this loss trains more quickly without clear loss of embedding quality. This loss operates on elements of the cross-correlation matrix, $\mathcal{C}$, computed between the vector embeddings of each modality. The hyperparameter λ controls the relative weight of on- and off- diagonal cross-correlations.

$$\mathcal{C}_{ij} = \frac{\sum_b z_{b,i}^S z_{b,j}^P}{\sqrt{\sum_b (z_{b,i}^S)^2}\sqrt{\sum_b (z_{b,i}^P)^2}} \quad (2)$$

$$\mathcal{L}_{\text{Barlow}} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2 \quad (3)$$

Here, b indexes batch samples, i, j index vector dimensions, and S and P identify SMILES and point embeddings, respectively.

3.2.3 Aggregate Loss

The model is trained with the sum of $\mathcal{L}_{\text{contrastive}}$ and an autoregressive entropy loss with $\mathcal{L}_{\text{contrastive}} \in \mathcal{L}_{\text{infoNCE}}, \mathcal{L}_{\text{Barlow}}$ weighted by another hyperparameter, β.

$$\mathcal{L}_{AR} = -\langle \log(P(x_i | x_{j<i})) \rangle_{\text{minibatch}} \quad (4)$$

$$\mathcal{L} = \mathcal{L}_{\text{AR}} + \beta \mathcal{L}_{\text{contrastive}} \quad (5)$$

This scheme is well-suited to rapidly screening large text-based corpora of small molecules without any preprocessing, but there are shortcomings which could be addressed in future work. In particular, because multiple SMILES strings can be made for one molecule, the autoregressive loss unfairly penalizes valid and desired generations. We experimented with an auxiliary loss term that incorporated molecular property information (See Appendix 8.1.1) but ultimately found that it reduced training stability.

3.3 Dataset

The training set for our contrastive learning model consists of more than 140 million (SMILES, geometry) tuples aggregated from several sources: ChEMBL (Gaulton et al., 2012), GEOM-Drugs (Axelrod and Gómez-Bombarelli, 2022), TensorMol (Yao et al., 2017), Mcule compounds (`https://mcule.com/database/`), ZINC22 (Tingle et al., 2023), and 54M combinatorial molecules enumerated from Enamine's catalog of building

blocks (`https://enamine.net/building-blocks/building-blocks-catalog`). The 3D coordinates are generated by several different processes. In the case of GEOM-Drugs and TensorMol, geometries are the result of DFT calculations, and in all other cases they are RDKit conformers optimized by MMFF94s. The mixed quality of geometries is intentional to increase the generality of our point encoder. The TensorMol geometries feature many structures which are significantly distorted from equilibrium. Train/test/validation splitting is achieved with fixed ranges of md5-hashes of the SMILES string. Some summary statistics of the conformer distribution of the dataset are given in Appendix 8.3. Exploring the precise effects of conformer ensemble quality on COATI results is left as an interesting avenue for future work.

3.4 Training

Table 2 contains hyperparameters of the COATI model variants we experimented with during this work. All networks are implemented in PyTorch (Paszke et al., 2019) and optimized with AdamW (Kingma and Ba, 2014), with $\beta_1, \beta_2 = \{0.9, 0.99\}$. Models were trained in float precision with a cosine learning rate schedule starting at $5{*}10^{-4}$. Experiments with mixed precision (bfloat16) for the transformer head led to worse losses and were not pursued further. Dropout is not used throughout, but a weight decay of $\eta = 0.1$ is applied. Models were trained using 16 A100 80GB GPUs, distributed across two DGX nodes each. Models were trained to approximate convergence, typically after 7 billion tokens had been iterated through during training. Held-out validity and isomorphism statistics were monitored in order to determine convergence.

As the total space of hyperparameters is large and training a single model is fairly expensive, we focused our effort and resources on 256-dimensional models which is the maximum embedding size which is well-behaved with this point encoder architecture. Attempts to increase the point embedding dimension beyond 256 suffer from initialization or smoothing issues. Additionally, we train a model (Autoreg_Only) that does *not* use any contrastive information in order to determine how much value the contrastive loss adds over a SMILES transformer. There is significant evidence that the limited expressive power of the point encoder relative to the transformer head is a major limitation of the model as-developed. This sets up a compromise between regression performance (which is driven by representation independence and strong contrastive loss against the point representation) and autoencoding (which is driven by the transformer). This is a clear direction for improvement in future work.

We note that batch size does not behave as a COATI hyperparameter in the same way it does in networks which do not couple training examples. In the limit of single-row batches (which is often used to fine-tune large language models with batch gradient aggregation), the contrastive losses would never contain any contrastive information. To accelerate training, we run models at the largest batch size possible, and find anecdotally that large batch sizes provide the most stable training. In practice, this leads to batch sizes of $\sim$2048 for 256-dimensional models across all GPUs.

During training, we randomly apply the following augmentations to training data: injecting a token with the molecule's dataset of origin (see Section 3.3 for datasets), randomizing the order of SMILES strings as in Arús-Pous et al. (2019), and permuting substring order following the fill-in-the-middle procedure of Bavarian et al. (2022). See Appendix 8.4 for examples of augmented strings.

PyTorch code of the model, exploratory notebooks, and trained checkpoints are open source and available at `https://github.com/terraytherapeutics/COATI/`.

4 Results

4.1 Generation and Autoencoding

COATI models can be used to perform molecular generation and autoencoding by encoding (from either a SMILES string or a molecular point cloud) to produce a latent vector describing molecular identity, and decoding by injecting the latent vector into the transformer using the procedure described in Section 3.2. Unless otherwise mentioned, generations are performed using GPT-2's top-k scheme with k=100 and an inverse temperature of 2 (Radford et al., 2019), although we find in practice that fidelity and validity of autoencoding and generation are not sensitive to these choices. We observe that the best COATI variants achieve generation validity upwards of 98%, and autoencoding ability on par with benchmark models that use descriptors to condition the encoder/decoder scheme (Winter et al., 2019). We also observe, interestingly, that a SMILES transformer model trained without contrastive information performs very well at this autoencoding task, but is unable to perform multi-modal encoding. We provide a table containing results of all COATI variants in Table 3 of Appendix 8.5. We interrogate generation/encoding failure modes in the next section using decoding likelihood as a proxy measure.

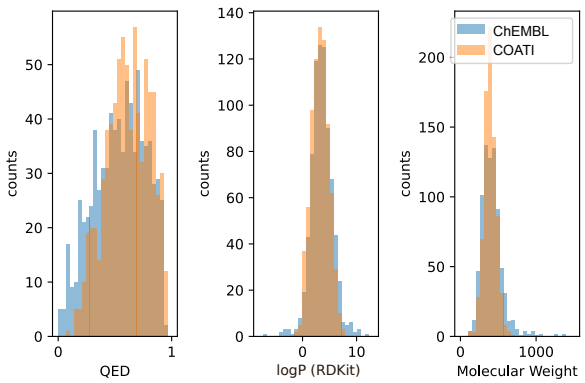

Figure 2: Conditional sampling from the space of ChEMBL-like molecules produces molecules with druglike properties. These histograms contain RDKit property distributions of ChEMBL samples overlaid with COATI samples from the set of ChEMBL molecules.

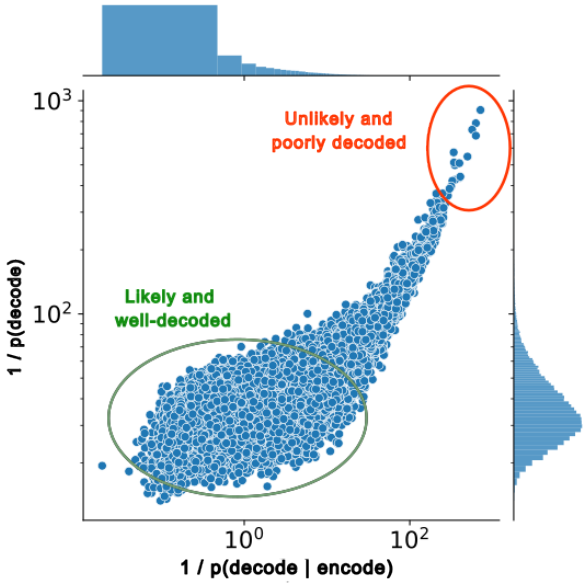

Figure 3: Decoding errors tend to occur in rare molecules. Negative log likelihoods of generating ChEMBL samples with the embedding token are plotted versus the unconditional negative log-likelihood of the same molecule. This indicates that poorly autoencoded molecules are *highly unlikely* - and that perhaps autoencoding performance can be improved by adding unlikely molecules to training.

4.2 Chemical Space Generation

The use of tokens encoding a molecule's dataset of origin (described in Section 3.4) allows us to prompt the autoregressive transformers to generate samples not related to a specific molecule, but from a *set* of molecules as designated during training. Figure 2 shows property histograms computed from molecules decoded via the prompt `[SET][chembl_mols][SMILES]`. These properties are not encoded or decoded by our model. We see that the distributions of the quantitative estimate of druglikeness (QED), lipophilicity, and molecular weight are close matches to the "real" ground-truth distribution of ChEMBL molecules, indicating that the statistical properties of the molecule set have been learned by the transformer. One could easily imagine other molecular properties encoded as tokens, suitable for conditioned generation.

The mean likelihood of generating druglike molecules with and without embedding is an important metric of any generative model, and both the mean and distributions of both quantities for a sample from ChEMBL are given in Figure 3. Most of the molecules in the low-likelihood tail have very high molecular weight, and are somewhat unfairly represented by comparing likelihoods of vastly different molecular weights. We also wanted to examine if the likelihood of generating a molecule without a conditioning embedding token indicates how likely the molecule is to decode given the embedding token. This has ramifications for use of the embedding space in a generative model, because the likelihood of unconditioned generation is determined by the composition of the training set. Indeed, we find that the two likelihoods are closely related (Figure 3). Many large-scale generative models of chemical space, including this work, use available catalogs of molecules as training data and are appropriate generators of rapidly accessible chemical space. However, this raises questions about the generative likelihood of synthetically accessible but unavailable, and physically stable but synthetically challenging molecules. We have no practical purpose to pursue generation of "likely unavailable" molecules at this time, but prospective users should bear this limitation in mind.

4.3 Regression with Linear Probes

A critical feature of a molecular representation is its ability to be "decoded" not only to molecular structure but to useful molecular properties, such as potency against a drug target or Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) properties that are critical to a molecule's pipeline progression. Unfortunately, almost all publicly available realistic regression datasets for chemical tasks are small, typically under 1000 data points; therefore, we have created an ensemble of tasks with data from diverse sources, augmenting publicly available data with millions of data points from the Terray platform. In this experiment, we compare frozen embeddings from COATI (text and point), CDDD (Winter et al., 2019), ChemBERTa MTR (Ahmad et al., 2022), ChemGPT (Frey et al., 2022), CLAMP (Seidl et al., 2023), MegaMolBART (mmb, 2022) as well as fingerprints from 2048-dimensional ECFP6 (ECFP6 2048) (Rogers and Hahn, 2010), RDKit fingerprints (RDKit FP), and RDKit 2D normalized descriptors (Kelley, 2023) on real-world activity regression tasks. To avoid advantaging or disadvantaging embeddings downstream based on the expressive power of a regressor, we fit a linear regressor (Pedregosa et al., 2011) for each embedding. The performance of each embedding for a given task is scaled from 0 to 1 - proportional to the best linear regressor for that task such that a score of 1 means the embedding was the best for that task.

To provide a more practical evaluation of molecule-target binding data, we leverage data from the Terray platform described in Section 5.2. This assay produces millions of scalar values correlating with the binding affinity of a molecule to a protein target. We consider four proteins for the molecule binding affinity datasets: Bruton's Tyrosine Kinase (BTK), Human Carbonic Anhydrase II (Sly and Hu, 1995) (hCAII), Protein-1, and Protein-2, whose names we are unable to release. Each protein has a training and test dataset of binding measurements, as well as a held-out set of biochemical activity assay data. We evaluate embeddings on their predictive power for the binding test split by sensitivity of the model at distinguishing the top-2% of binding molecules. We also evaluate the models on rank correlation to the held out set of biochemical assay measurements for BTK, Protein-1, and Protein-2 (hCAII activity results not available).

COATI learned representations outperform or match other learned representations when ordered by performance across all tasks (Figure 4). We see that 2048-dimensional ECFP6 fingerprints perform very well on the binding tasks (Table 4) - indicating that variation in the data is well-explained by graph structural features. On the small ADMET datasets, RDKit 2D normalized descriptors perform well but occasionally seem to overfit as do other representations.

During ablation studies (see full results in Table 4), we discovered that an autoregression-only SMILES transformer *underperforms* relative to contrastively trained COATI models, but other SMILES language models like CDDD (Winter et al., 2019) perform competitively. We hypothesize that the additional supervision in the form of properties provides a useful advantage, although combining contrastive training with additional property supervision (the FP variants) did not improve regression results. Without directly training on properties, COATI representations still perform competitively to models that used property information during training (Figure 4).

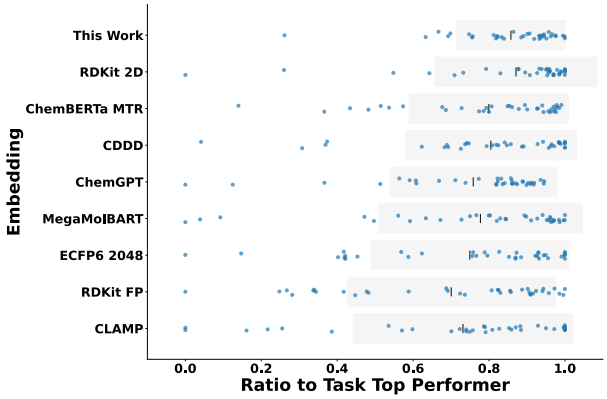

Figure 4: Embedding performance using a linear regressor for ADMET and binding tasks. Results are scaled proportionally to the best performance for a given task to better assess the performance across all tasks. Embeddings are ordered by the mean of their relative performances divided by their standard deviation. The highlighted COATI model (This Work) was trained using Barlow loss and inferred via SMILES input (Barlow_Closed).

We also see no discernible trends between the point representation (i.e., leveraging the 3D encoder to encode a conformer) and SMILES-based representations. This is certainly interesting - it is possible that 3D information does not contribute well to this task, the conformers (generated via RDKit MMFF94s) are too low quality to have any meaningful signal, or the training process of the 3D encoder focused on removing 3D information for the sake of contrastive optimization. We look forward to further exploring the properties of 3D encoders in a self-supervised setting in future work.

We evaluated the performance of linear regressors (described above) trained using frozen COATI embeddings on a set of ADMET datasets from the Therapeutic Data Commons (TDC) (Huang et al., 2021) and MoleculeNet (Wu et al., 2017; Ramsundar et al., 2019). These tasks represent typical noisy and data-limited chemical problems. Information on the datasets and training approach can be found in the Table 5. COATI embedding models perform consistently well when compared to the best linear regressor for each task. While it might be expected that featurizations trained or augmented with molecule properties would perform best on these datasets, COATI models without a mechanism for property reinforcement

perform comparably to similar learned representations and traditional methods.

5 Conditional Generation of Therapeutics

5.1 Background

A common task in early-stage drug discovery is often referred to as *hit-to-lead optimization*, where molecular design teams leverage information from a high-throughput screen to propose new molecules with improved potency or other properties. We assume a realistic molecule design scenario where one would like to make a generative model from a small ($\sim<$ 1M) number of experimental samples of a desired molecular property, possibly simultaneously optimizing or constraining several properties at once (Fromer and Coley, 2022). For example, it is often desirable to improve potency while constraining lipophilicity, due to its relationship to both protein-ligand affinity and various ADME properties (Johnson et al., 2018).

One approach to generation conditioned on these properties would be to fine-tune the generative model with a small number of samples as in Blaschke et al. (2020), although in our view this is more costly and cumbersome than using the frozen pre-trained embedding. We propose a method that achieves a "separation of concerns" - leveraging the continuous nature of the COATI latent space to train differentiable regressors using the pre-trained embedding, optimize properties directly on the vector space of the embedding, and decode to obtain molecules with desired properties (Anstine and Isayev, 2023; Bilodeau et al., 2022). This also allows us to easily use regression methodologies which quantify uncertainty.

5.2 Terray Platform Data

To build differentiable regressors and optimize compounds, we used an experimental dataset of 1,307,908 (combinatorial molecule, binding affinity) data points collected using the Terray platform. To briefly summarize the assay: molecules are produced by combinatorial synthesis tethered to polymer beads immobilized in an ultradense microarray on a silicon chip. A fluorescently-labeled target protein is flowed over the chip, incubated, and microscopy is used to quantify the amount of bound protein as a ratio of fluorescence on each bead to an empty background. The target protein considered in this section is Human Carbonic Anhydrase 2 (Sly and Hu, 1995), for which several hundred pIC50 data points are also available in ChEMBL (Mendez et al., 2018) (which we make use of as independent test data). None of the combinatorial molecules from the Terray dataset occur in the ChEMBL data, nor are the potency data collected from the same type of assay. Potency rank-ordering on the Terray platform corresponds well with low-throughput standard measurements of IC50 (Lebakken et al., 2009).

5.3 Differentiable Regressor with COATI Embeddings

These (combinatorial molecule, binding affinity) data points were embedded with COATI (Barlow_Closed), and the resulting 256D embedding vectors were used as training data for a DUE regressor (van Amersfoort et al., 2022). This regressor returns a normal distribution $(\mu_{\text{binding}}(v), \sigma_{\text{binding}}(v))$ as a function of a COATI vector v with an uncertainty estimate proportional to the input vector's distance to the training data . The outsample Pearson correlation between the log fluorescence ratio inferred by the COATI-DUE model with Terray platform data and experimental pIC50 tabulated in ChEMBL is 23%. We also trained DUE regressors for other chemical properties: RDKit-determined QED (Bickerton et al., 2012) and logP. These regressors were trained on a dataset of a few million molecules from the COATI dataset.

5.4 Metadynamics Generative Algorithm

One can easily use COATI to draw molecules randomly, using the conditional generation method explored in Section 4.2, and filter them for desired properties with these regressors. However, there are a few reasons to try more sophisticated optimizations. Binding potency is extremely sparse in chemical space, i.e., most druglike molecules do not bind to most targets. In a realistic regression model, trained with a realistic amount of data, potency is also sparse over vector space and riddled with local maxima making random sampling inefficient. Instead, we seek to treat the problem of selecting desirable compounds as a differentiable optimization of a vector-valued function. This allows us to exploit the smoothness of our learned chemical space, and focus on activity basins with room to chemically modify the lead.

We perform gradient ascent to maximize inferred potency in the DUE model, utilising Lagrange multipliers ($\lambda_{\text{QED}}, \lambda_{\text{logP}}$) to enforce druglikeness constraints (QED $>$ .5 and logP < 5). A rate of $2 * 10^{-3}$ is used in the gradient ascent, and molecules are decoded from the optimization vector every 15 iterations. We have found that optimizing the $\mu - \sigma$ of the DUE potency model is sufficient to keep the optimization from straying into a space of molecules and vectors too far from the training data. This objective function is somewhat analogous to other methods that utilize the predicted mean and variance of a Bayesian model (Srinivas et al., 2012) to balance exploration and exploitation, although we utilize the gradient of our model directly (along with other constraints specific to differentiable functions).

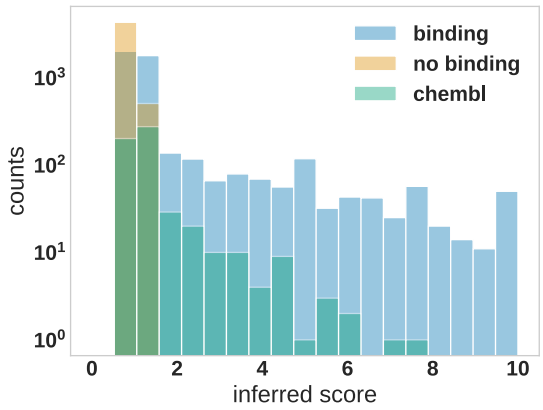

Figure 5: Optimizing binding affinity via metadynamics enriches generated molecules for predicted binding. Histograms above show binding scores for molecules generated during optimization of potency with constraints on QED and logP ("binding"), ChEMBL samples ("chembl"), and optimization purely for QED with a constraint on logP ("no binding").

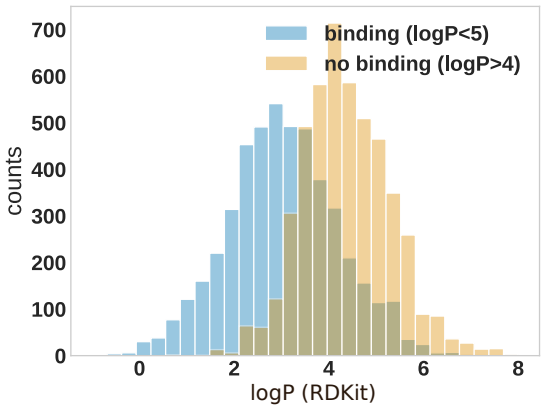

Figure 6: Lagrangian penalties are obeyed during optimization. Adding a Lagrangian penalty to binding optimization produces a distribution of molecules (blue) that largely satisfies the desired constraint of logP < 5.

We find that this optimization will rapidly reach local maxima, and so, inspired by the idea of metadynamics (Laio and Parrinello, 2002), we add a 256D isotropic Gaussian bump of standard deviation 0.125 periodically to the potential if the decoded molecule is unchanged after 25 steps. The final objective function minimized for 40,000 steps was:

$$\mathcal{L}(\vec{v}, \lambda_{\text{QED}}, \lambda_{\text{logP}}) = -(\mu_{\text{bind}}(v) - \sigma_{\text{bind}}(v)) + \lambda_{\text{logP}}((\mu_{\text{logP}}(v) - 5)^{+})^2 + \lambda_{\text{QED}}((0.5 - \mu_{\text{QED}}(v))^{+})^2 + \mathcal{L}_{\text{bump}}(v) \quad (6)$$

The results of this generative optimization task are summarized in Figures 5-6. The data labeled as "binding" refer to the primary optimization task – maximizing binding affinity while constraining QED and logP. As a control experiment (labeled "no binding" in Figures 5-6), and also to further demonstrate constraints, three independent trajectories were drawn from an objective which optimizes QED instead of potency. Instead of constraining logP < 5, as we do in our potency optimization experiments, it constrains logP > 4, which is difficult to satisfy since logP < 5 is correlated with high QED. Relative to filtering random samples, optimization is especially useful with constraints. Due to the nature of gradient descent, and the action of the added Gaussian bumps, the Lagrangian constraints are not perfectly satisfied over the entire trajectory, but are generally well-satisfied (Figure 6). We find that 94.6% of the potency trajectory samples satisfy the logP < 5 constraint, and 63% of the QED samples satisfy the logP > 4 constraint.

5.5 Realistic Generations from Experimental Data

As a realistic test of this metadynamics algorithm and learned COATI representation against a related method, we ran five trajectories from five randomly drawn molecules using COATI + Metadynamics and CDDD + QMO (Hoffman et al., 2022), another gradient-based generative approach to multi-objective molecule optimization. The starting points, regression method, and training data (consisting only of Terray platform data) of the two methods are the same, and only the encoder-decoder and gradient algorithm differ. The results of these generative experiments are summarized in Figure 7, along with the pIC50 data from ChEMBL. The COATI + Metadynamics method travels close to potent molecules known from ChEMBL, while the CDDD + QMO trajectories find fewer molecules near the ChEMBL molecules based on Tanimoto similarity. As a best-effort "impartial score" of binding affinity, we trained a separate ECFP6 regressor to score the unique generations from both methods. This provides a sense of whether or not each optimization method is exploiting artifacts of their embedding space. The COATI + Metadynamics generative optimization procedure succeeds in achieving high similarity to known potent molecules, despite the fact that these molecules do not occur in the training data. A movie of molecules visited in an example optimization process is available in the source code repository at `https://github.com/terraytherapeutics/COATI/blob/main/examples/binding_meta.gif`. Further details of this methodology and the experimental setup can be found in Appendix 8.9.

6 Conclusions

We have presented COATI, a contrastive framework for training decodable multi-modal molecular encoders. We provide several variants of these models, and show that

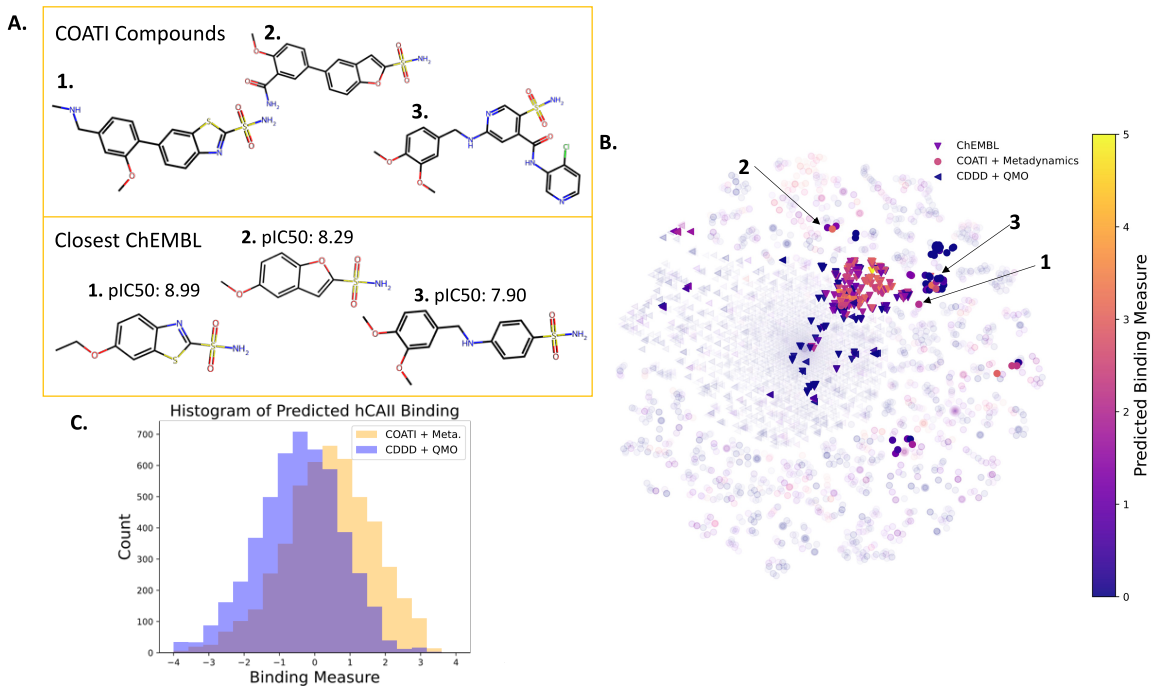

Figure 7: Metadynamics generation in the COATI latent space produces druglike molecules similar to known hits from ChEMBL. A) Examples of compounds found by COATI + Metadynamics and each of their nearest neighbors in ChEMBL by Tanimoto similarity. B) A visualization of COATI + Metadynamics along with CDDD + QMO. The t-SNE contains five metadynamics trajectories, five QMO trajectories, and out of sample ChEMBL hCAII data. Points with a Tanimoto similarity to ChEMBL < .3 are faded. Embeddings are generated from ECFP6 fingerprints and scored by a DUE model trained on ECFP6 representations. Compounds from (A) are identified by black arrows. C) A histogram of predicted binding for all trajectories based on the same ECFP6 DUE used in (B). Note that this ECFP6 model was not used during optimization.

they can perform generation/autoencoding from a latent vector space into textual molecular representations. We have shown that the learned embedding is usefully expressive and produces linear models which match or outperform commonly used fingerprints for a real-world, large-scale molecule binding assay and low-data ADMET datasets, without task-specific feature engineering. We further contribute a metadynamics-inspired molecular design algorithm that leverages a unique set of high-quality data to perform practical, constrained molecular optimization.

We found that while it is possible to produce a fixed-length representation that is both decodable and useful for molecular property regression, there exists some tension between these two goals. We demonstrate that focusing on validity of autoencoding over a large dataset can lead to poor regression performance. Several representations including this work have sought to constrain the embedding to also represent molecular properties which may be physically connected to downstream tasks in different ways. Our work is premised on the hypothesis that 3D structure would produce a more smooth and continuous embedding of chemical space than autoencoded text alone. Other chemical representation models also leverage supplementary sources of information during pre-training, such as pre-computed physiochemical heuristic features used by Winter et al. (2019), or textual descriptions of assays (Seidl et al., 2023). We find that pre-training additional properties improves robustness versus ECFP6 fingerprints (see Section 4.3), but underperforms on binding affinity tasks. The paucity of publicly available data at the present time makes the relative merits of pre-training choices somewhat unclear, and supports the need for larger public datasets to support development of molecular encoders.

We demonstrated our end-to-end training scheme with two contrastive losses and two common encoders, but there is nothing preventing the approach from being applied to new point encoders, textual representations, or contrastive loss functions. Highly expressive molecular graph encoders have been an important area of research activity for decades, and it is certain that, especially for 3D representations, there are encoders more powerful than the E(3)-GNN used in this work. A clear and useful direction for future work will be conditioned generation of molecular conformers on the basis of this latent space, or latent diffusion to produce molecules with desired properties while remaining close to an embedded molecule. We see COATI models and their successors as providing a path towards a unified representation of molecular structures and conformations that can be used productively for many tasks.

The field of generative molecular design and optimization is very new and promising for accelerating the process of bringing effective therapeutics into the clinic. It will likely progress differently than the meteoric rise of image or text data, simply because accurate, practically useful data is far more difficult to acquire. In our view, coupling generative design with novel assay technologies provides an attractive practical advantage over filtering or virtual screening, which we have demonstrated via constrained generative optimization using realistic, large-scale data from the Terray platform.

7 Acknowledgements

The authors thank Zahid Panjwani and Alex Hesselgrave for infrastructure support, and Alice Liu, Jacob Berlin, and Eli Berlin for comments on the manuscript. We would like to thank Nvidia Corporation for use of DGX Cloud resources for model training. We would also like to thank the Terray library, decoding, screening, and automation teams for their work producing platform data.

References

Jean-Louis Reymond. The chemical space project. *Accounts of Chemical Research*, 48(3):722–730, 2015. doi:10.1021/ar500432k. URL https://doi.org/10.1021/ar500432k. PMID: 25687211.

Oleksandr O Grygorenko, Dmytro S Radchenko, Igor Dziuba, Alexander Chuprina, Kateryna E Gubina, and Yurii S Moroz. Generating multibillion chemical space of readily accessible screening compounds. *iScience*, 23(11):101681, Nov 2020. ISSN 2589-0042 (Electronic); 2589-0042 (Linking). doi:10.1016/j.isci.2020.101681.

Yan Xu. Introduction of galaxi – the wuxi apptec virtual space, 2021. URL https://www.biosolveit.de/wp-content/uploads/2021/08/Xu.pdf. PowerPoint presentation.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Lingle Wang, Yujie Wu, Yuqing Deng, Byungchan Kim, Levi Pierce, Goran Krilov, Dmitry Lupyan, Shaughnessy Robinson, Markus K. Dahlgren, Jeremy Greenwood, Donna L. Romero, Craig Masse, Jennifer L. Knight, Thomas Steinbrecher, Thijs Beuming, Wolfgang Damm, Ed Harder, Woody Sherman, Mark Brewer, Ron Wester, Mark Murcko, Leah Frye, Ramy Farid, Teng Lin, David L. Mobley, William L. Jorgensen, Bruce J. Berne, Richard A. Friesner, and Robert Abel. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137(7):2695–2703, 2015. doi:10.1021/ja512751q. URL https://doi.org/10.1021/ja512751q. PMID: 25625324.

Christoph Bannwarth, Sebastian Ehlert, and Stefan Grimme. Gfn2-xtb—an 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. doi:10.1021/acs.jctc.8b01176. URL https://doi.org/10.1021/acs.jctc.8b01176. PMID: 30741547.

Danny Reidenbach, Micha Livne, Rajesh K. Ilango, Michelle Gill, and Johnny Israeli. Improving small molecule generation using mutual information machine, 2023.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. Enhancing activity prediction models in drug discovery with the ability to understand human language, 2023.

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. doi:10.1021/acs.jcim.0c00915. URL https://doi.org/10.1021/acs.jcim.0c00915. PMID: 33118816.

Robin Winter, Floriane Montanari, Frank Noé, and Djork-Arné Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.*, 10:1692–1701, 2019. doi:10.1039/C8SC04175J. URL http://dx.doi.org/10.1039/C8SC04175J.

Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation, 2021. URL https://arxiv.org/abs/2106.04399.

Chengyi Liu, Wenqi Fan, Yunqing Liu, Jiatong Li, Hang Li, Hui Liu, Jiliang Tang, and Qing Li. Generative diffusion models on graphs: Methods and applications, 2023.

Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation, 2022. URL https://arxiv.org/abs/2209.14734.

Pedro O. Pinheiro, Joshua Rackers, Joseph Kleinhenz, Michael Maser, Omar Mahmood, Andrew Martin Watkins, Stephen Ra, Vishnu Sresht, and Saeed Saremi. 3d molecule generation by denoising voxel grids, 2023.

Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, and Xing Xie. Graphformers: Gnn-nested transformers for representation learning on textual graph, 2021. URL https://arxiv.org/abs/2105.02605.

Jinwoo Kim, Tien Dat Nguyen, Seonwoo Min, Sungjun Cho, Moontae Lee, Honglak Lee, and Seunghoon Hong. Pure transformers are powerful graph learners, 2022. URL https://arxiv.org/abs/2207.02505.

Dominic Masters, Josef Dean, Kerstin Klaser, Zhiyi Li, Sam Maddrell-Mander, Adam Sanders, Hatem Helal, Deniz Beker, Ladislav Rampášek, and Dominique Beaini. Gps++: An optimised hybrid mpnn/transformer for molecular property prediction, 2022. URL https://arxiv.org/abs/2212.02229.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021a. URL https://arxiv.org/abs/2103.00020.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 3d infomax improves gnns for molecular property prediction. 2021. doi:10.48550/ARXIV.2110.04126. URL https://arxiv.org/abs/2110.04126.

Yihao Xue, Kyle Whitecross, and Baharan Mirzasoleiman. Investigating why contrastive learning benefits robustness against label noise. In *International Conference on Machine Learning*, pages 24851–24871. PMLR, 2022.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. URL https://arxiv.org/abs/2102.12092.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. URL https://arxiv.org/abs/2112.10752.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL https://arxiv.org/abs/2205.11487.

Geemi P. Wellawatte, Aditi Seshadri, and Andrew D. White. Model agnostic generation of counterfactual explanations for molecules. *Chem. Sci.*, 13:3697–3705, 2022. doi:10.1039/D1SC05259D. URL http://dx.doi.org/10.1039/D1SC05259D.

Smiles - a simplified chemical language. URL https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html.

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, oct 2020. doi:10.1088/2632-2153/aba947. URL https://doi.org/10.1088%2F2632-2153%2Faba947.

Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C. Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, Rafael F. Lameiro, Dominik Lemm, Alston Lo, Seyed Mohamad Moosavi, José Manuel Nápoles-Duarte, AkshatKumar Nigam, Robert Pollice, Kohulan Rajan, Ulrich Schatzschneider, Philippe Schwaller, Marta Skreta, Berend Smit, Felix Strieth-Kalthoff, Chong Sun, Gary Tom, Guido Falk von Rudorff, Andrew Wang, Andrew D. White, Adamo Young, Rose Yu, and Alán Aspuru-Guzik. Selfies and the future of molecular string representations. *Patterns*, 3(10):100588, 2022. ISSN 2666-3899. doi:https://doi.org/10.1016/j.patter.2022.100588. URL https://www.sciencedirect.com/science/article/pii/S2666389922002069.

Shion Honda, Shoi Shi, and Hiroki R. Ueda. SMILES transformer: Pre-trained molecular fingerprint for low data drug discovery. *CoRR*, abs/1911.04738, 2019. URL http://arxiv.org/abs/1911.04738.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48, 2017. doi:10.1186/s13321-017-0235-x. URL https://doi.org/10.1186/s13321-017-0235-x.

Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi:10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t. PMID: 20426451.

Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, 2002. doi:10.1021/ci010132r. URL https://doi.org/10.1021/ci010132r. PMID: 12444722.

Raymond E Carhart, Dennis H Smith, and R Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73, 1985.

Alice Capecchi, Daniel Probst, and Jean-Louis Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12(1):43, 2020. doi:10.1186/s13321-020-00445-4. URL https://doi.org/10.1186/s13321-020-00445-4.

Eugene N. Muratov, Jürgen Bajorath, Robert P. Sheridan, Igor V. Tetko, Dmitry Filimonov, Vladimir Poroikov, Tudor I. Oprea, Igor I. Baskin, Alexandre Varnek, Adrian Roitberg, Olexandr Isayev, Stefano Curtalolo, Denis Fourches, Yoram Cohen, Alan Aspuru-Guzik, David A. Winkler, Dimitris Agrafiotis, Artem Cherkasov, and Alexander Tropsha. Qsar without borders. *Chem. Soc. Rev.*, 49:3525–3564, 2020. doi:10.1039/D0CS00098A. URL http://dx.doi.org/10.1039/D0CS00098A.

Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jürgen Bajorath. Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8):3186–3204, 2014. doi:10.1021/jm401411z. URL https://doi.org/10.1021/jm401411z. PMID: 24151987.

Ingo Muegge and Prasenjit Mukherjee. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin Drug Discov*, 11(2):137–148, 2016. ISSN 1746-045X (Electronic); 1746-0441 (Linking). doi:10.1517/17460441.2016.1117070.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry, 2017. URL https://arxiv.org/abs/1704.01212.

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. doi:10.1021/acscentsci.7b00572. URL https://doi.org/10.1021/acscentsci.7b00572. PMID: 29532027.

Wengong Jin, Regina Barzilay, and Tommi S. Jaakkola. Junction tree variational autoencoder for molecular graph generation. *CoRR*, abs/1802.04364, 2018. URL http://arxiv.org/abs/1802.04364.

Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. 2018. doi:molgan. URL https://arxiv.org/abs/1805.11973.

J. S. Smith, O. Isayev, and A. E. Roitberg. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chem. Sci.*, 8:3192–3203, 2017. doi:10.1039/C6SC05720A. URL http://dx.doi.org/10.1039/C6SC05720A.

Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8(1):13890, 2017. doi:10.1038/ncomms13890. URL https://doi.org/10.1038/ncomms13890.

Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. 2017. doi:10.48550/ARXIV.1706.08566. URL https://arxiv.org/abs/1706.08566.

Kun Yao, John E. Herr, David W. Toth, Ryker Mcintyre, and John Parkhill. The tensormol-0.1 model chemistry: a neural network augmented with long-range physics, 2017. URL https://arxiv.org/abs/1711.06385.

Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances*, 7(15):eabe4166, 2021. doi:10.1126/sciadv.abe4166. URL https://www.science.org/doi/abs/10.1126/sciadv.abe4166.

Kevin McCloskey, Eric A. Sigel, Steven Kearnes, Ling Xue, Xia Tian, Dennis Moccia, Diana Gikunju, Sana Bazzaz, Betty Chan, Matthew A. Clark, John W. Cuozzo, Marie-Aude Guié, John P. Guilinger, Christelle Huguet, Christopher D. Hupp, Anthony D. Keefe, Christopher J. Mulhern, Ying Zhang, and Patrick Riley. Machine learning on dna-encoded libraries: A new paradigm for hit finding. *Journal of Medicinal Chemistry*, 63(16):8857–8866, 2020. doi:10.1021/acs.jmedchem.0c00452. URL https://doi.org/10.1021/acs.jmedchem.0c00452. PMID: 32525674.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds, 2018a. URL https://arxiv.org/abs/1802.08219.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks, 2020. URL https://arxiv.org/abs/2006.10503.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks, 2021a. URL https://arxiv.org/abs/2102.09844.

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. doi:10.1038/s41598-019-47148-x. URL https://doi.org/10.1038/s41598-019-47148-x.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models, 2022. URL https://arxiv.org/abs/2209.01712.

Gayane Chilingaryan, Hovhannes Tamoyan, Ani Tevosyan, Nelly Babayan, Lusine Khondkaryan, Karen Hambardzumyan, Zaven Navoyan, Hrant Khachatrian, and Armen Aghajanyan. Bartsmiles: Generative masked language models for molecular representations, 2022. URL https://arxiv.org/abs/2211.16349.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12): 1256–1264, 2022. doi:10.1038/s42256-022-00580-7. URL https://doi.org/10.1038/s42256-022-00580-7.

Ingoo Lee and Hojung Nam. Infusing linguistic knowledge of smiles into chemical language models, 2022. URL https://arxiv.org/abs/2205.00084.

Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. Graphaf: a flow-based autoregressive model for molecular graph generation. *arXiv preprint arXiv:2001.09382*, 2020.

Victor Garcia Satorras, Emiel Hoogeboom, Fabian B. Fuchs, Ingmar Posner, and Max Welling. E(n) equivariant normalizing flows, 2021b. URL https://arxiv.org/abs/2105.09016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021b. URL https://arxiv.org/abs/2103.00020.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021a.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations, 2020. URL https://arxiv.org/abs/2010.13902.

Dong Chen, Kaifu Gao, Duc Duy Nguyen, Xin Chen, Yi Jiang, Guo-Wei Wei, and Feng Pan. Algebraic graph-assisted bidirectional transformers for molecular property prediction. *Nature Communications*, 12(1):3521, 2021. doi:10.1038/s41467-021-23720-w. URL https://doi.org/10.1038/s41467-021-23720-w.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry, 2021. URL https://arxiv.org/abs/2110.07728.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *CoRR*, abs/2103.03230, 2021b. URL https://arxiv.org/abs/2103.03230.

Jinhua Zhu, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Unified 2d and 3d pre-training of molecular representations, 2022. URL https://arxiv.org/abs/2207.08806.

2022. URL https://github.com/NVIDIA/MegaMolBART.

Nathan Frey, Ryan Soklaski, Simon Axelrod, Siddharth Samsi, Rafael Gómez-Bombarelli, Connor Coley, and Vijay Gadepally. Neural scaling of deep chemical models. 05 2022. doi:10.26434/chemrxiv-2022-3s512.

Arne Schneuing, Yuanqi Du, Charles Harris, Arian Jamasb, Ilia Igashov, Weitao Du, Tom Blundell, Pietro Lió, Carla Gomes, Max Welling, Michael Bronstein, and Bruno Correia. Structure-based drug design with equivariant diffusion models, 2023.

Oliver Unke, Mihail Bogojeski, Michael Gastegger, Mario Geiger, Tess Smidt, and Klaus-Robert Müller. Se(3)-equivariant prediction of molecular wavefunctions and electronic densities. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 14434–14447. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/78f1893678afbeaa90b1fa01b9cfb860-Paper.pdf.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds, 2018b. URL https://arxiv.org/abs/1802.08219.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields, 2022.

M.A. Osipov, B.T. Pickup, and D.A. Dunmur. A new twist to molecular chirality: intrinsic chirality indices. *Molecular Physics*, 84(6):1193–1206, 1995. doi:10.1080/00268979500100831. URL https://doi.org/10.1080/00268979500100831.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021. URL https://arxiv.org/abs/2104.09864.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Ron Mokady, Amir Hertz, and Amit H. Bermano. Clipcap: Clip prefix for image captioning, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P Overington. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*, 40(Database issue):D1100–7, Jan 2012. ISSN 1362-4962 (Electronic); 0305-1048 (Print); 0305-1048 (Linking). doi:10.1093/nar/gkr777.

Simon Axelrod and Rafael Gómez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022. doi:10.1038/s41597-022-01288-4. URL https://doi.org/10.1038/s41597-022-01288-4.

Benjamin I. Tingle, Khanh G. Tang, Mar Castanon, John J. Gutierrez, Munkhzul Khurelbaatar, Chinzorig Dandarchuluun, Yurii S. Moroz, and John J. Irwin. Zinc-22-a free multi-billion-scale database of tangible compounds for ligand discovery. *Journal of Chemical Information and Modeling*, 63(4):1166–1176, 2023. doi:10.1021/acs.jcim.2c01253. URL https://doi.org/10.1021/acs.jcim.2c01253. PMID: 36790087.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

Josep Arús-Pous, Simon Viet Johansson, Oleksii Prykhodko, Esben Jannik Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. Randomized smiles strings improve the quality of molecular generative models. *Journal of Cheminformatics*, 11(1):71, 2019. doi:10.1186/s13321-019-0393-0. URL https://doi.org/10.1186/s13321-019-0393-0.

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. Efficient training of language models to fill in the middle, 2022. URL https://arxiv.org/abs/2207.14255.

Brian Kelley. Descriptastorus: Python descriptor creator and utilities. https://github.com/bp-kelley/descriptastorus, 2023.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

William S. Sly and Peiyi Y. Hu. Human carbonic anhydrases and carbonic anhydrase deficiencies. *Annual Review of Biochemistry*, 64(1):375–401, 1995. doi:10.1146/annurev.bi.64.070195.002111. URL https://doi.org/10.1146/annurev.bi.64.070195.002111. PMID: 7574487.

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.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: A benchmark for molecular machine learning. 2017.

Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. *Deep Learning for the Life Sciences*. O'Reilly Media, 2019. https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

Jenna C. Fromer and Connor W. Coley. Computer-aided multi-objective optimization in small molecule discovery, 2022.

Ted W. Johnson, Rebecca A. Gallego, and Martin P. Edwards. Lipophilic efficiency as an important metric in drug design. *Journal of Medicinal Chemistry*, 61(15):6401–6420, 2018. doi:10.1021/acs.jmedchem.8b00077. URL https://doi.org/10.1021/acs.jmedchem.8b00077. PMID: 29589935.

Dylan M. Anstine and Olexandr Isayev. Generative models as an emerging paradigm in the chemical sciences. *Journal of the American Chemical Society*, 145(16):8736–8750, 04 2023. doi:10.1021/jacs.2c13467. URL https://doi.org/10.1021/jacs.2c13467.

Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F. Jensen. Generative models for molecular discovery: Recent advances and challenges. *WIREs Computational Molecular Science*, 12(5):e1608, 2022. doi:https://doi.org/10.1002/wcms.1608. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1608.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R Leach. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, 11 2018. ISSN 0305-1048. doi:10.1093/nar/gky1075. URL https://doi.org/10.1093/nar/gky1075.

Connie S Lebakken, Steven M Riddle, Upinder Singh, W Jack Frazee, Hildegard C Eliason, Yi Gao, Laurie J Reichling, Bryan D Marks, and Kurt W Vogel. Development and applications of a broad-coverage, tr-fret-based kinase binding assay platform. *Journal of biomolecular screening*, 14(8):924–935, 2009.

Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. On feature collapse and deep kernel learning for single forward pass uncertainty, 2022.

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. doi:10.1038/nchem.1243. URL https://doi.org/10.1038/nchem.1243.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, may 2012. doi:10.1109/tit.2011.2182033. URL https://doi.org/10.1109%2Ftit.2011.2182033.

Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Science*, 99(20):12562–12566, October 2002. doi:10.1073/pnas.202427399.

Samuel C. Hoffman, Vijil Chenthamarakshan, Kahini Wadhawan, Pin-Yu Chen, and Payel Das. Optimizing molecules using efficient queries from property evaluations. *Nature Machine Intelligence*, 4(1):21–31, 2022. doi:10.1038/s42256-021-00422-y. URL https://doi.org/10.1038/s42256-021-00422-y.

Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1):20, 2015. doi:10.1186/s13321-015-0069-3. URL https://doi.org/10.1186/s13321-015-0069-3.

Pasquale Foggia, Carlo Sansone, and Mario Vento. An improved algorithm for matching large graphs. 01 2001.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

8 Appendices

8.1 Descriptor Decoding Auxiliary Loss

8.1.1 Descriptor Decoding

We experimented with adding an auxiliary loss that is designed to encourage encoding of additional chemical modalities, although this variant is not the preferred checkpoint discussed in the results. We extended the model to linearly map the `[STOP]` token vector from our SMILES encoder to 2048-dimensional ECFP6 fingerprints (Rogers and Hahn, 2010) and 200-dim RDKit 2D normalized descriptors (Kelley, 2023). These were selected given their strong empirical performance on the regression tasks evaluated in the results. The additional loss $\mathcal{L}_{FP}$ is the sum of binary cross-entropy loss over Morgan bits and mean squared error loss over RDKit 2D normalized descriptors:

$$\mathcal{L}_m = \frac{1}{2048} \sum_i y_i \log(\sigma(f_m(z_s)_i)) + (1 - y_i) \log(1 - \sigma(f_m(z_s)_i)) \tag{7}$$

$$\mathcal{L}_{2d} = \frac{1}{200} \sum_i (y_i - f_{2d}(z_s)_i)^2 \tag{8}$$

$$\mathcal{L}_{FP} = w(\mathcal{L}_m + \mathcal{L}_{2d}) \tag{9}$$

where w is a weight hyperparameter. This loss was added to the aggregate loss described in Section 3.2.3 and applied batchwise.

We found that this loss decreased training stability and produced suboptimal results.

8.2 Model Variants

Model	Loss	Vocab	Aux. Loss	E(3)-GNN	Transformer	Latent Dim.
Tall_Closed	InfoNCE + AR	Closed	No	4*128	16*16*128	128
Grande_Closed	InfoNCE + AR	Closed	No	5*256	16*16*256	256
Grande_Closed_FP	InfoNCE + AR + FP	Closed	Yes	5*256	16*16*256	256
Barlow_Closed_FP	Barlow + AR + FP	Closed	Yes	5*256	16*16*256	256
Barlow_Closed	Barlow + AR	Closed	No	5*256	16*16*256	256
Autoreg_Only	AR	Closed	No	N/A	16*16*256	256
Barlow_Venti	Barlow + AR	Closed	No	5*256	16*16*384	384
Grande_Open	InfoNCE + AR	Open	No	5*256	16*16*256	256
SELFIES_Barlow	Barlow + AR	SELFIES	No	5*256	16*16*256	256

Table 2: COATI model variants. The E(3)-GNN and Transformer columns specify number of layers * hidden dim and number of layers * number of heads * hidden dim, respectively.

8.3 Dataset Conformers

GEOM-Drugs and TensorMol are sets of multiple DFT-optimized conformers per molecule of which there are 24.3 per molecule on average in GEOM-Drugs, and 20.1 per molecule on average in TensorMol. All other datasets have five initial conformers generated using RDKit's ETKDG implementation and optimized via MMFF94s, with the lowest energy conformer selected. SMILES strings were deliberately standardized differently for different datasets to provide a diversity of representations - for example, TensorMol SMILES strings contain explicit hydrogens while other datasets do not.

8.4 Tokenization

8.4.1 Vocab Generation

Tokenizer vocabulary was initialized with a simple set of characters required to encode SMILES strings, along with a set of "sentinel" tokens used to encode augmentations. We iteratively apply a BPE procedure (Radford et al., 2019) to training data, iteratively combining tokens from 400 batches of 256 entries. Tokens that appeared more than 1000 times were added to the vocabulary to produce the "open" vocabulary, which contains approximately 11,996 tokens. The closed vocabulary was produced by pruning the original vocabulary of tokens with unbalanced parenthesis (i.e., an incomplete branch of the graph). An additional run of token aggregation was performed after pruning, producing a total of 8,726 tokens.

Ablation	Validity	Canonical Match	Match (no Chiral)	Matching 2D Connectivity
Tall_Closed	.917 ± .020	.734 ± .029	.741 ± .028	.759 ± .026
Grande_Closed	.986 ± .007	.877 ± .027	.888 ± .026	.902 ± .026
Grande_Closed_FP	.981 ± .009	.885 ± .021	.897 ± .019	.911 ± .018
Barlow_Closed_FP	.945 ± .016	.785 ± .028	.792 ± .027	.812 ± .029
Barlow_Closed	.987 ± .008	.914 ± .020	.928 ± .018	.939 ± .019
Autoreg_Only	.973 ± .013	.980 ± .009	.981 ± .009	.981 ± .009
Barlow_Venti	.984 ± .009	.862 ± .028	.882 ± .025	.895 ± .024
Grande_Open	.902 ± .021	.860 ± .031	.867 ± .030	.885 ± .026
SELFIES_Barlow	1.000 ± 0	.748 ± .032	.759 ± .029	.782 ± .029
CDDD	.993 ± .005	.941 ± .015	.942 ± .014	.962 ± .012

Table 3: Generative Metrics: SMILES-to-SMILES. We see practically equivalent performance from closed-vocabulary models (with Barlow loss slightly outperforming infoNCE loss), and on par with the CDDD character-level baseline.

8.4.2 Augmentation Tokens

Several special tokens are reserved for implementing the data augmentations described in Section 3.2. Provided below are a few examples of (reconstructed) SMILES strings with special tokens included:

```
[EMBED][UNK][SMILES]CNC(=O)CN1CCCN(C(=O)c2cc(-n3cccc3)ccc2Br)CC1[STOP]
```

```
[EMBED][UNK][SMILES]CCOc1ccc(F)c(-c2cc(F)c[MIDDLE]([SUFFIX]C(=O)N3CCN(CCC#N)CC3)cc2C)c1[STOP]
```

```
[EMBED][UNK][SMILES]Cc1ccccc1NC(=O)C[MIDDLE]N1C(=O)N/C(=C\\c2cccn2-c2cccc(C(=O)O)[SUFFIX]c2)C1=O[SET][geom_drugs][STOP]
```

Note that in the original implementation, the `[EMBED]` tokens are listed as `[CLIP]`. The `[UNK]` tokens are used as the actual positional injection of the embedding vector.

8.5 Generation Metrics

We use several types of metrics to assess the quality of COATI for translation between molecular representations and generation. Tanimoto similarity (Bajusz et al., 2015) of ECFP4 fingerprints (Rogers and Hahn, 2010) will be invoked as a measure of molecular similarity. *Validity* of a SMILES string or graph are both binary measures determined by RDKit's valence rules. We evaluate *identity* using three variants, corresponding to different levels of "strictness" - bond graph isomorphism without atom types (Matching 2D Connectivity), bond graph isomorphism with atom types (Match (no Chiral)), and whether or not the canonical SMILES match exactly (Canonical Match). We compute isomorphism using the VF2 algorithm (Foggia et al., 2001) as implemented in NetworkX (Hagberg et al., 2008). These answer different questions about the representation, and in practice we find that the differences between the latter two metrics are driven by chirality and tautomerization. Unless otherwise mentioned, generations are performed using GPT-2's top-k scheme with k=100 and an inverse temperature of 2 (Radford et al., 2019), although we find in practice that the fidelity and validity metrics we report are not sensitive to these choices.

8.6 Autoencoding and Decodability

We assess the ability of COATI models to encode molecules and decode into SMILES strings. We focus on the "SMILES-to-SMILES" task for comparison to other methods, where a SMILES strings is encoded using the transformer module described in Section 3.1.2, and decoded using the same transformer. Table 3 contains summary statistics for the validity, similarity, and isomorphism metrics described in Section 8.5, across multiple variants of COATI models. We also evaluate the SMILES-to-SMILES decoding metrics on the RNN-based autoencoder from Winter et al. (2019), another fixed-size encoder-decoder model we refer to as CDDD. We note that CDDD's training process was designed to decode several molecular properties represented in our regression test set and that it was not trained on SMILES with stereochemistry.

We see a few interesting properties of the model variants - the first of which is that models only trained on (augmented) SMILES text sequences (Autoreg_Only and CDDD) perform excellently on this task, indicating that the contrastive training is not strictly required for the task of producing valid SMILES strings, or autoencoding the strings themselves. However, Figure 8 suggests that autoregressive loss is not sufficient for learning a well-behaved latent space as the Autoreg_Only models latent space does not interpolate between molecules smoothly. Autoregressive models with additional contrastive or molecular property losses do not have this same issue.

We also see the effect of tokenizer engineering on validity - the model variants with a closed-parenthetical vocabulary have an easier time producing valid SMILES strings versus similar models that allow for open-parenthesis tokens.

Table 4: Nominal test results for the linear regression suite. COATI model embeddings either come from the text representation of the molecule (often SMILES) or 3D points.

Task	Metric	Barlow_Closed (Text)	RDKit 2D	ChemBERTa MTR	CDDD	ChemGPT	MegaMolBART	ECFP6 2048	RDKit FP	CLAMP	Tall_Closed (Point)	Grande_Open (Point)	Tall_Closed (Text)	Grande_Closed (Point)	Barlow_Venti (Point)	Grande_Open (Text)	Barlow_Closed_FP (Text)	Barlow_Venti (Text)	Barlow_Closed_FP (Point)	Barlow_Closed (Point)	Grande_Closed (Text)	SELFIES_Barlow (Text)	Autoreg_Only
Ames Mutagenicity	AUROC	0.77	0.81	0.80	0.82	0.71	0.80	0.71	0.66	0.78	0.80	0.78	0.70	0.78	0.79	0.72	0.77	0.75	0.80	0.80	0.74	0.75	0.68
Bioavailability	AUROC	0.65	0.72	0.66	0.58	0.63	0.53	0.61	0.60	0.42	0.58	0.57	0.59	0.63	0.59	0.68	0.65	0.66	0.48	0.53	0.68	0.59	0.54
CYP P450 1A2 Inhib.	AUROC	0.90	0.91	0.91	0.92	0.87	0.92	0.87	0.87	0.92	0.89	0.90	0.88	0.89	0.89	0.90	0.91	0.90	0.90	0.90	0.89	0.90	0.84
CYP P450 2C19 Inhib.	AUROC	0.84	0.86	0.86	0.87	0.78	0.87	0.84	0.83	0.90	0.85	0.85	0.83	0.85	0.84	0.84	0.86	0.85	0.86	0.86	0.83	0.84	0.78
CYP P450 2C9 Inhib.	AUROC	0.85	0.86	0.86	0.73	0.81	0.86	0.84	0.80	0.90	0.85	0.86	0.84	0.86	0.85	0.85	0.86	0.85	0.86	0.86	0.84	0.85	0.80
CYP P450 2D6 Inhib.	AUROC	0.82	0.83	0.84	0.76	0.79	0.82	0.82	0.79	0.86	0.83	0.83	0.81	0.82	0.82	0.83	0.84	0.82	0.84	0.83	0.82	0.82	0.76
CYP P450 3A4 Inhib.	AUROC	0.83	0.84	0.85	0.86	0.78	0.85	0.82	0.80	0.87	0.83	0.85	0.82	0.85	0.84	0.82	0.85	0.83	0.85	0.85	0.82	0.81	0.77
Caco-2	RMSE	0.69	0.43	1.19	1.16	0.59	0.92	0.57	1.04	2.0	0.53	0.52	0.56	0.65	0.45	0.60	0.57	0.75	0.52	0.64	0.50	0.60	0.70
Clearance Hepatocyte	Spearman	0.27	0.36	0.28	0.28	0.27	0.24	0.23	0.24	-0.03	0.31	0.29	0.41	0.39	0.31	0.32	0.38	0.28	0.25	0.16	0.30	0.19	0.16
Clearance Microsome	Spearman	0.41	0.57	0.47	0.37	0.33	0.29	0.27	0.20	0.15	0.48	0.51	0.52	0.45	0.51	0.29	0.51	0.38	0.39	0.50	0.59	0.39	0.30
ClinTox	AUROC	0.88	0.86	0.78	0.85	0.73	0.79	0.77	0.61	0.62	0.77	0.67	0.83	0.76	0.77	0.81	0.83	0.76	0.81	0.76	0.78	0.80	0.71
DILI	AUROC	0.68	0.74	0.77	0.62	0.78	0.69	0.83	0.86	0.76	0.90	0.83	0.77	0.75	0.76	0.66	0.62	0.51	0.74	0.80	0.64	0.67	0.61
HIA	AUROC	0.92	0.93	0.90	0.83	0.88	0.82	0.96	0.88	0.86	0.88	0.93	0.91	0.86	0.96	0.91	0.87	0.81	0.87	0.80	0.84	0.89	0.83
Half Life	RMSE	91.6	92.1	170	578	65.3	616	57.3	71.0	148	71.0	101	67.6	83.8	90.0	99.8	111	157	97.4	107	96.1	94.1	80.1
LD50	RMSE	0.98	0.97	0.95	0.98	0.97	1.01	1.16	1.26	0.91	0.95	0.97	1.04	0.96	0.98	1.02	0.95	1.02	1.02	0.92	1.01	0.95	1.00
Lipophilicity	RMSE	0.97	0.87	0.84	0.79	1.11	0.84	1.27	1.77	0.98	0.96	0.91	1.02	0.99	0.93	1.00	0.84	0.93	0.86	0.90	0.99	0.99	1.15
PAMPA Permeability	AUROC	0.71	0.73	0.73	0.70	0.70	0.73	0.61	0.68	0.70	0.77	0.77	0.74	0.77	0.68	0.72	0.77	0.69	0.72	0.73	0.73	0.75	0.68
PPBR	RMSE	15.3	13.5	14.8	14.9	15.3	15.7	29.8	47.0	16.6	12.6	14.3	14.3	13.6	14.5	15.4	14.1	16.3	14.6	13.6	15.9	15.0	17.0
Pgp Inhibition	AUROC	0.87	0.90	0.80	0.87	0.84	0.84	0.81	0.79	0.73	0.93	0.88	0.87	0.84	0.87	0.87	0.88	0.82	0.90	0.84	0.86	0.86	0.80
Solubility, AqSolDB	RMSE	1.25	1.08	1.15	1.08	1.32	1.14	1.84	2.24	1.26	1.28	1.19	1.29	1.21	1.2	1.2	1.18	1.25	1.32	1.21	1.21	1.28	1.56
Volume of Distribution	Spearman	0.30	0.29	0.14	0.11	-0.03	-0.02	0.04	0.07	-0.02	0.17	0.19	0.19	0.22	0.09	0.14	0.20	0.18	0.16	0.23	-0.02	0.21	0.10
hERG Blockers	AUROC	0.74	0.79	0.65	0.60	0.71	0.53	0.71	0.61	0.72	0.83	0.60	0.81	0.68	0.70	0.68	0.70	0.62	0.66	0.78	0.75	0.79	0.68
hERG Central Inhib.	AUROC	0.83	0.84	0.85	0.85	0.79	0.86	0.86	0.85	0.86	0.83	0.83	0.81	0.84	0.84	0.82	0.84	0.83	0.84	0.84	0.82	0.82	0.75
hERG, Karim et al.	AUROC	0.80	0.82	0.83	0.84	0.72	0.84	0.85	0.85	0.79	0.80	0.81	0.76	0.81	0.80	0.78	0.81	0.80	0.82	0.82	0.79	0.78	0.73
BACE Classification	AUROC	0.76	0.74	0.75	0.61	0.75	0.82	0.63	0.57	0.65	0.82	0.78	0.79	0.81	0.75	0.81	0.73	0.76	0.82	0.83	0.77	0.79	0.83
BACE Regression	RMSE	0.81	3295	1.38	2.56	0.95	1.17	1.88	2.32	1.05	0.81	0.95	0.89	0.85	0.90	0.83	0.96	0.90	0.88	0.88	0.86	0.84	0.92
HIV	AUROC	0.76	0.73	0.73	0.75	0.71	0.76	0.73	0.76	0.92	0.76	0.76	0.78	0.77	0.75	0.79	0.78	0.77	0.74	0.76	0.77	0.76	0.75
Solubility, Delaney	RMSE	0.62	0.54	1.19	0.75	0.63	0.64	1.23	1.84	0.86	0.57	0.62	0.54	0.63	0.56	0.60	0.52	0.63	0.63	0.68	0.55	0.55	0.79
Tox21	AUROC	0.75	0.79	0.77	0.77	0.74	0.76	0.69	0.71	0.76	0.76	0.76	0.75	0.77	0.76	0.73	0.75	0.76	0.77	0.76	0.76	0.76	0.70
BTK Activity	Spearman	0.20	0.22	0.11	0.18	0.19	0.02	-0.10	-0.00	0.08	0.18	0.13	0.18	0.15	0.16	0.21	0.17	0.14	0.09	0.10	0.21	0.14	0.15
BTK Binding	Sensitivity	0.35	0.42	0.41	0.40	0.25	0.42	0.39	0.35	0.35	0.34	0.37	0.29	0.36	0.36	0.33	0.40	0.33	0.36	0.37	0.31	0.28	0.23
Protein-1 Activity	Spearman	0.37	0.36	0.34	0.50	0.30	0.28	0.50	0.24	0.36	0.40	0.40	0.34	0.42	0.44	0.41	0.28	0.37	0.30	0.33	0.40	0.23	0.36
Protein-1 Binding	Sensitivity	0.34	0.27	0.35	0.36	0.25	0.41	0.48	0.46	0.26	0.30	0.35	0.26	0.34	0.35	0.32	0.34	0.36	0.34	0.32	0.36	0.32	0.25
Protein-2 Activity	Spearman	0.22	0.19	0.16	0.22	0.04	0.22	0.12	0.29	0.22	0.29	0.19	0.12	0.09	0.19	0.21	0.13	0.18	0.18	0.12	0.20	0.07	0.03
Protein-2 Binding	Sensitivity	0.39	0.35	0.37	0.39	0.29	0.39	0.48	0.46	0.35	0.32	0.36	0.31	0.37	0.36	0.33	0.36	0.37	0.38	0.40	0.33	0.33	0.30
hCAII Binding	Sensitivity	0.66	0.58	0.54	0.65	0.55	0.64	0.62	0.60	0.52	0.66	0.64	0.56	0.61	0.63	0.59	0.63	0.59	0.59	0.64	0.59	0.57	0.49

Experiments early in development studying the weight of contrastive and autoregressive losses revealed a competition between point-contrastive loss, and autoregressive loss, which is related to validity. We found that downweighting off-diagonal (i.e., explicitly contrastive) terms of the loss function (both Barlow and InfoNCE) improved training speed and stability. We hypothesize that some of this tension is due to the limited expressive power of the point encoder vs. the text encoder.

8.7 Conformational Degeneracy and Isomerization

Atomic positions in molecules fluctuate significantly at any temperature, and so the mapping between 3D conformations and 2D bond representations is many-to-one. The learned shared representation must map different conformations onto the same vector, throwing away conformation information. However, if atoms move enough to constitute a change in bonding, the two arrangements are isomers (different molecules with the same numbers of atoms), which should map to different COATI representations.

2D embeddings of COATI representations from multiple conformations of four closely related probe molecules (dimethyl fumarate, fumaric acid, ribose, and glucose) and a set of their isomers (C6H8O4, C4H4O4, C5H10O5, and C6H12O6, respectively) can be viewed in Figure 9. An additional 10 chemical formulas containing multiple conformations of approximately 10 training set isomers each were also added to expand the 2D embedding space. Sets of isomers are distinguishable and related by similarity. Isomers of glucose and ribose are shown to be related in the embedding space as would be expected. The same expectation for dimethyl fumarate and fumaric acid is also demonstrated.

8.8 Regression Suite Results

Nominal values that were used to construct the scaled values in Figure 4 can be seen in Table 4. Descriptions of the public ADMET datasets can be found in Table 5 while binding and activity datasets are characterized in Section 5.2

8.9 hCAII Optimization Overview

For experimental setup, five molecules were used as starting points by randomly sampling from the COATI holdout set and were used as the starting point for trajectories for COATI + Metadynamics and CDDD + QMO. Each trajectory was

Table 5: Overview of ADMET datasets

Dataset	Source	Type	Size	Metric	Unit	Split	Description
Ames Mutagenicity	TDC	Class.	7255	AUROC	Binary	Scaffold	The Ames test is a short-term bacterial reverse mutation assay detecting a large number of compounds which can induce genetic damage and frameshift mutations. The dataset is aggregated from four papers
Bioavailability	TDC	Class.	640	AUROC	Binary	Scaffold	Given a drug SMILES string, predict the activity of bioavailability.
CYP P450 1A2 Inhib.	TDC	Class.	12579	AUROC	Binary	Scaffold	The CYP P450 genes are involved in the formation and breakdown (metabolism) of various molecules and chemicals within cells. Specifically, CYP1A2 localizes to the endoplasmic reticulum.
CYP P450 2C19 Inhib.	TDC	Class.	12665	AUROC	Binary	Scaffold	CYP2C19 gene provides instructions for making an enzyme called the endoplasmic reticulum, which is involved in protein processing and transport.
CYP P450 2C9 Inhib.	TDC	Class.	12092	AUROC	Binary	Scaffold	The CYP P450 2C9 plays a major role in the oxidation of both xenobiotic and endogenous compounds.
CYP P450 2D6 Inhib.	TDC	Class.	13130	AUROC	Binary	Scaffold	The CYP P450 genes are involved in the formation and breakdown of various molecules and chemicals within cells. Specifically, CYP2D6 is primarily expressed in the liver and in areas of the central nervous system.
CYP P450 3A4 Inhib.	TDC	Class.	12328	AUROC	Binary	Scaffold	CYP3A4 is an important enzyme in the body, mainly found in the liver and in the intestine.
Caco-2	TDC	Reg.	906	RMSE	log(cm/s)	Scaffold	The experimental result on the rate of drug passing through the human colon epithelial cancer cell line can approximate the rate at which the drug permeates through the human intestinal tissue
Clearance Hepatocyte	TDC	Reg.	1020	Spearman	uL/min per million cells	Scaffold	The volume of plasma cleared of a drug over a specified time period. This is a dataset curated from ChEMBL database containing experimental results on intrinsic clearance, deposited from AstraZeneca.
Clearance Microsome	TDC	Reg.	1102	Spearman	uL/min/g	Scaffold	The volume of plasma cleared of a drug over a specified time period. This is a dataset curated from ChEMBL database containing experimental results on intrinsic clearance, deposited from AstraZeneca.
ClinTox	TDC	Class.	1484	AUROC	Binary	Scaffold	The ClinTox dataset includes drugs that have failed clinical trials for toxicity reasons and also drugs that are associated with successful trials.
DILI	TDC	Class.	475	AUROC	Binary	Scaffold	Drug-induced liver injury (DILI) is fatal liver disease caused by drugs and it has been the single most frequent cause of safety-related drug marketing withdrawals for the past 50 years (e.g. iproniazid, ticrynafen, benoxaprofen). This dataset is aggregated from U.S. FDA's National Center for Toxicological Research
HIA	TDC	Class.	578	AUROC	Binary	Scaffold	The ability of a drug to be absorbed from the human gastrointestinal system into the bloodstream of the human body
Half Life	TDC	Reg.	667	Spearman	hr	Scaffold	The duration for the concentration of the drug in the body to be reduced by half. Obtained from ChEMBL assay 1614674
LD50	TDC	Reg.	7385	RMSE	log(kg/mol)	Scaffold	The conservative dose that can lead to lethal adverse effects.
Lipophilicity	TDC	Reg.	4200	RMSE	log-ratio	Scaffold	Lipophilicity measures the ability of a drug to dissolve in a lipid (e.g. fats, oils) environment. From MoleculeNet
PAMPA Permeability	TDC	Class.	2035	AUROC	Binary	Scaffold	PAMPA (parallel artificial membrane permeability assay) is a commonly employed assay to evaluate drug permeability across the cellular membrane. PAMPA does not model active and efflux transporters but the majority of drugs are absorbed by passive diffusion through the membrane
PPBR	TDC	Reg.	1797	RMSE	Binary	Scaffold	The human plasma protein binding rate (PPBR) is expressed as the percentage of a drug bound to plasma proteins in the blood. From a ChEMBL assay deposited by AstraZeneca
Pgp Inhibition	TDC	Class.	1212	AUROC	Binary	Scaffold	P-glycoprotein (Pgp) is an ABC transporter protein involved in intestinal absorption, drug metabolism, and brain penetration, and its inhibition can seriously alter a drug's bioavailability and safety.
Solubility, AqSolDB	TDC	Reg.	9982	RMSE	log(mol/L)	Scaffold	Aqeuous solubility measures a drug's ability to dissolve in water.
Volume of Distribution	TDC	Reg.	1130	Spearman	L/kg	Scaffold	The volume of distribution at steady state (VDss) measures the degree of a drug's concentration in body tissue compared to concentration in blood.
hERG Blockers	TDC	Class.	648	AUROC	Binary	Scaffold	Human ether-a-go-go related gene (hERG) is crucial for the coordination of the heart's beating.
hERG Central Inhib.	TDC	Class.	306893	AUROC	Binary	Scaffold	Given a drug, predict whether it blocks hERG with an IC50 <10uM.
hERG, Karim et al.	TDC	Class.	13445	AUROC	Binary	Scaffold	A integrated Ether-a-go-go-related gene (hERG) dataset consisting of molecular structures labelled as hERG (<10uM) and non-hERG (>=10uM) blockers in the form of SMILES strings was obtained from the DeepHIT, the BindingDB database, ChEMBL bioactivity database, and other literature
BACE Classification	MoleculeNet	Class.	1513	AUROC	Binary	Scaffold	Provides bindings results for a set of inhibitors of human beta-secretase (BACE-1)
BACE Regression	MoleculeNet	Reg.	1513	RMSE	pIC50	Scaffold	Provides bindings results for a set of inhibitors of human beta-secretase (BACE-1)
HIV	MoleculeNet	Class.	40000	AUROC	Binary	Scaffold	A dataset which tested the ability to inhibit HIV replication
Solubility, Delaney	MoleculeNet	Reg.	1128	RMSE	log(mol/L)	Scaffold	A regression dataset containing structures and water solubility data
Tox21	MoleculeNet	Class.	8000	AUROC	Binary	Random	The Toxicology in the 21st Century (Tox21) initiative created a public database measuring the toxicity of compounds

run to maximize a measure of hCAII binding predicted by a DUE Regressor with constraints that QED > .5 and logP < 5.

For COATI + Metadynamics the objective function took the form of Equation 9 in the main text and was minimized over 40,000 steps per trajectory with molecules being decoded every 25 steps. For CDDD + QMO, steps are taken using a zeroth-order gradient descent method outlined in Hoffman et al. (2022). The minimized objective function took the form:

$$\mathcal{L}(\vec{v}) = -(\mu_{bind}(v) - \sigma_{bind}(v)) + \max(.5 - \text{QED}(\text{CDDD}(v)), 0) + \max(\text{logP}(\text{CDDD}(v)) - 5, 0) \quad (10)$$

where μ, σ are computed by the CDDD binding DUE regressor and $\text{CDDD}(v)$ is the molecule decoded from v by CDDD. In this case, QED and logP are computed on decoded molecules directly via RDKit. With references to the variables used in the QMO paper for optimization parameters, the starting learning rate (α_0) was .2, the number of samples to compute the pseudo-gradient (Q) was 150, and the pseudo-gradient smoothing parameter (β) was 10. Each trajectory was run for 1000 steps, as each step is more computationally expensive given the additional sampling required.

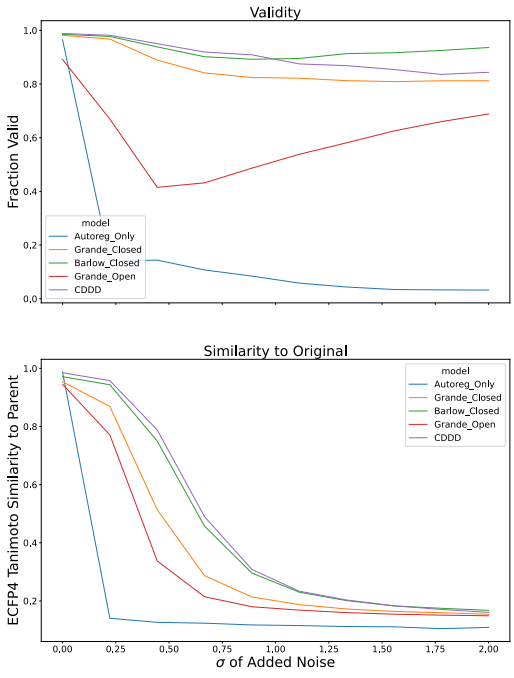

Figure 8: Similarity to reference compound and validity with added $\mathcal{N}(0, \sigma)$ noise over a random sample of 2500 molecules from the COATI datasets test partition. Three generation attempts were made for each molecule in the input.

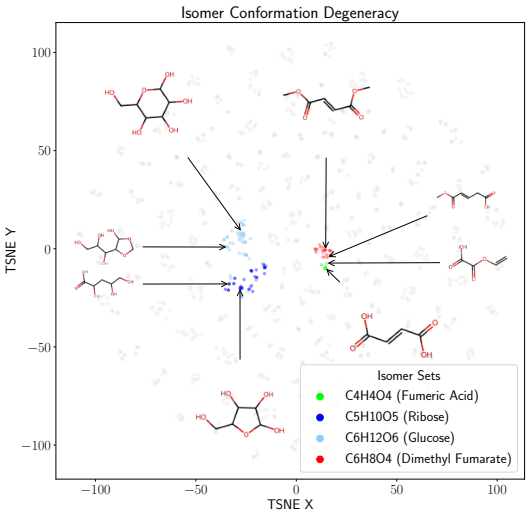

Figure 9: t-SNE of the COATI embeddings from the 3D point cloud of multiple conformations of each molecule in different sets of isomers. Many sets of isomers from the training data were added to fully express the space.