Crystal-GFN: sampling crystals with desirable properties and constraints

Mila AI4Science*
Mila, Quebec AI Institute

Alex Hernandez-Garcia
Mila, Quebec AI Institute
Université de Montréal
alex.hernandez-garcia@mila.quebec

Alexandre Duval
Mila, Quebec AI Institute
CentraleSupélec, Université Paris-Saclay
alexandre.duval@mila.quebec

Alexandra Volokhova
Mila, Quebec AI Institute
Université de Montréal
alexandra.volokhova@mila.quebec

Yoshua Bengio
Mila, Quebec AI Institute
Université de Montréal
yoshua.bengio@mila.quebec

Divya Sharma
Johns Hopkins University
dsharm23@jh.edu

Pierre Luc Carrier
Mila, Quebec AI Institute
pierre.luc.carrier@mila.quebec

Michał Koziarski
Mila, Quebec AI Institute
Université de Montréal
michal.koziarski@mila.quebec

Victor Schmidt
Mila, Quebec AI Institute
Université de Montréal
schmidtv@mila.quebec

Abstract

Accelerating material discovery holds the potential to greatly help mitigate the climate crisis. Discovering new solid-state materials such as electrocatalysts, superionic conductors or photovoltaic materials can have a crucial impact, for instance, in improving the efficiency of renewable energy production and storage. In this paper, we introduce Crystal-GFN, a generative model of crystal structures that sequentially samples structural properties of crystalline materials, namely the space group, composition and lattice parameters. This domain-inspired approach enables the flexible incorporation of physical and structural constraints, as well as the use of any available predictive model of a desired physico-chemical property as an objective function. To design stable materials, one must target the candidates with the lowest formation energy, which is used as an objective to evaluate the capabilities of Crystal-GFN. The formation energy of a crystal structure is predicted here by a new proxy model trained on MatBench. The results demonstrate that Crystal-GFN is able to sample diverse crystals with low formation energy.

*This team name is the way the authors have found to express that the customary ordered list of authors hardly reflects their contributions. All authors have actively and significantly contributed to this work. The list in this manuscript has been randomised. Correspondence email: alex.hernandez-garcia@mila.quebec

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

1 Introduction

Materials discovery plays a vital role in transforming numerous industries that are currently responsible for a significant fraction of the global greenhouse gas emissions. From developing new-generation photovoltaic panels or solid-state batteries to tackling carbon capture or catalysis, the quest for innovative materials with targeted properties has the power to reshape the technological landscape.

However, the discovery of new materials is an extremely complicated task at various levels of innovation: crystal structure generation, property prediction and developing successful synthesis protocols. In practice, before scientists can think of designing a material with a desirable property, they must search through the vast space of possible materials and use their domain expertise to investigate a few/ handful of promising structures. This traditional trial-and-error process is extremely costly both in terms of time and efforts, providing machine learning (ML) with an opportunity to greatly accelerate the generation and evaluation of promising candidates.

While ML has already made significant progress with respect to property prediction and generation of small molecules [14, 19, 17, 25][2], its impact in the field of solid-state materials lags behind. At the root of this discrepancy lies the particular structure of materials, whose periodicity in 3D space imposes periodic boundary conditions that make it difficult to use finite graphs for crystal representation. As a result, the sole task of generating a stable candidate is complex. It involves exploring the combinatorial space of chemical compositions, considering all possible periodic arrangements of atoms as well as respecting complex physics that make them viable, stable and metastable structures.

In this work, we introduce Crystal-GFN (Fig. 1), a generative model designed to sequentially sample crystal structures in a space inspired by theoretical crystallography. Instead of directly generating atomic positions, Crystal-GFN samples the space group, the composition and the lattice parameters of a structure (Section 3). This domain-inspired approach facilitates the incorporation of chemical and structural constraints, as well as the possibility to extend the sampling space, for example to obtain atomic positions. The GFlowNet framework offers the ability to sample proportionally to a reward function, which enables diverse sampling. Through an empirical evaluation (Section 4) of a Crystal-GFN trained with a predictive model of the formation energy, we demonstrate that our proposed method is able to sample diverse candidates with low formation energy.

2 Background

In this section, we briefly review the necessary background on GFlowNets and crystallography.

2.1 GFlowNets

Generative Flow Networks [2] (GFlowNets or GFN for short), were introduced as an amortised inference method to sample from high-dimensional distributions, where both traditional methods such as MC-MC and reinforcement learning are inefficient in terms of mode mixing. In essence,

[2]We refer the reader to Appendix A for a more detailed related work overview.

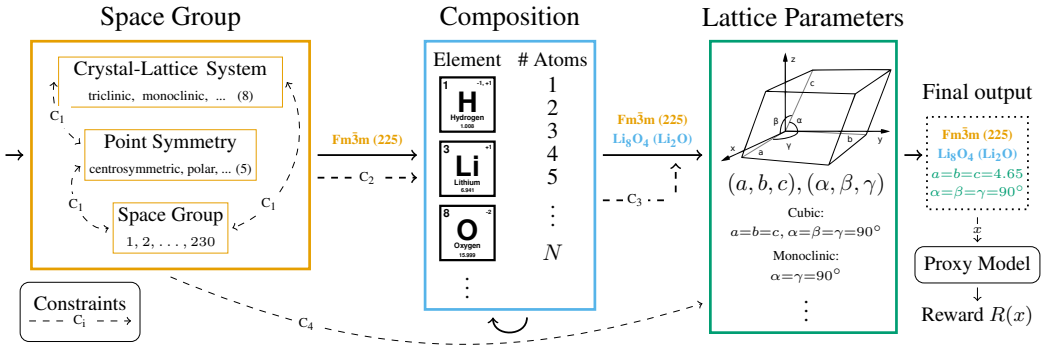


Figure 1: A schematic of the crystal generation process of Crystal-GFN.

GFlowNets are designed to learn a sampling policy π to generate objects $x \in \mathcal{X}$ proportionally to a non-negative reward function $R(x)$, that is $\pi(x) \propto R(x)$. This facilitates the discovery of multiple modes of a given function, which is a desirable objective in scientific discovery [12]. For instance, GFlowNets have been successfully applied for biological sequence design and molecular property prediction as part of active learning loops [11, 9].

A key property of GFlowNets is that objects are generated sequentially. Starting from a special state s_0, transitions $s_t \to s_{t+1} \in \mathbb{A}$ are applied between states $s \in \mathcal{S}$, forming trajectories $\tau = (s_0 \to s_1 \to \ldots \to x)$, where $\mathbb{A}$ is a predefined action space and $\mathcal{S}$ the state space. This sequential construction of objects, together with the fact that the transition policy is parameterised by a neural network (with parameters θ), provides amortisation and the potential of systematic generalisation. Here, we make use of the Trajectory Balance objective [15], which has proven effective in various tasks [11, 9, 15].

2.2 Crystals

Crystals are highly structured solid materials defined by a unit cell made of atom types and coordinates that repeat in the three spatial directions, as specified by the crystal's periodic lattice. This ordered structure gives them their unique shapes and properties and their understanding is crucial in the development of new materials. While there exists an infinite number of crystal structures, the field of theoretical crystallography has developed ways to systematically parameterise and classify crystals.

Lattice An n-dimensional *lattice* Λ can be defined as the set of integral combinations of the linearly independent *lattice basis vectors* $\mathbf{a}_i \in \mathbb{R}^n$: $\Lambda \doteq \{\sum_i^n m_i \mathbf{a}_i \mid m_i \in \mathbb{Z}\}$. In 3D, the 3 lattice basis vectors can be converted into 6 parameters: a, b, c determine the lengths of each dimension and α, β, γ the angles between dimensions.

Composition We call composition the set of all atom types present in the crystal's unit cell and their respective quantities. Together with the lattice Λ, the atomic positions provide a complete description of the crystal structure.

Space group A space group is the symmetry group of the crystal's repeating pattern in space, whose elements are the rigid transformations of the pattern that leave it unchanged. A space group element can be described as a tuple $(\mathbf{W}, \mathbf{t})$, where $\mathbf{W}$ is the linear part of the transformation and $\mathbf{t}$ a translation. An element maps a vector $\mathbf{x} \in \mathbb{R}^n$ to $\mathbf{W}\mathbf{x} + \mathbf{t}$. In three dimensions, space groups are classified into 230 types.

3 Crystal-GFNs

In this section, we describe the details of Crystal-GFN. Drawing inspiration from crystallography, we propose to represent crystals as the concatenation of three distinct components, or *subspaces*: space group (SG), composition (C) and lattice parameters (LP) of the unit cell of a crystal. The union of the spaces of these three components would make the *state space* $\mathcal{S} = \mathcal{S}_{SG} \times \mathcal{S}_C \times \mathcal{S}_{LP}$ and *sample space* $\mathcal{X} = \mathcal{X}_{SG} \times \mathcal{X}_C \times \mathcal{X}_{LP}$ of a naive, unrestricted implementation with GFlowNets. An important advantage of the GFlowNet framework and of the domain-inspired data representation that we propose to use is that it allows us to flexibly introduce domain knowledge such as well-studied geometrical constraints from crystallography as well as chemical constraints. These constraints are mostly introduced by restricting the action space $\mathbb{A}$ of the Crystal-GFN, both inter- and intra-subspace. We design the Crystal-GFN such that trajectories first select the space group, then the composition and finally the lattice parameters. Below we describe the three subspaces and the constraints we have introduced.

3.1 Space group

The sample space of the space group subspace of our Crystal-GFN is $\mathcal{X}_{SG} = \{1, 2, \ldots 230\}$, corresponding to the space of the 230 symmetry groups that exist in three dimensions. As discussed in Section 2.1, GFlowNets rely on the decomposition of objects into multiple steps to facilitate generalisation. Therefore, we draw inspiration from theoretical crystallography to incorporate additional structure into the space group subspace.

Crystals are classified into seven *crystal systems*: triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal and cubic. Each crystal system comprises a set of point groups. Another related but slightly different category is the lattice system, with also seven systems (see Fig. 3), each comprising a set of Bravais lattices. The lattice system category is convenient for our purposes because it imposes specific constraints on the lattice parameters. Here, we use a derived categorisation resulting from the combination of a crystal system and a lattice system. We refer to it as *crystal-lattice system* and it has eight categories: triclinic, monoclinic, orthorhombic, tetragonal, trigonal-rhombohedral, trigonal-hexagonal, hexagonal-hexagonal and cubic.

Besides the crystal-lattice system, we introduce another category in the Crystal-GFN space group subspace: the *point symmetry* (also known as site symmetry), which defines the type of symmetry of a point group. We use the following five categories of point symmetries: centrosymmetric, non-centrosymmetric, enantiomorphic, polar and enantiomorphic-polar. We choose these categories because the combination of point symmetry and crystal system gives rise to one of the 32 crystal classes or crystallographic point groups.

In sum, our space group subspace $\mathcal{S}_{SG}$ is a three-dimensional set where the entries correspond to the crystal-lattice system, the point symmetry and the space group. We define the action space such that, starting from the source state s_0, the transitions can set any of the three dimensions. However, the selection of an option along any dimension restricts the valid options for the remaining dimensions. For example, if point symmetry "non-centrosymmetric" is selected from the source state, only three crystal-lattice systems remain valid and the number of valid space groups is reduced to 25. The "stop" action is valid if and only if the space group has been selected.

3.2 Composition

We represent a composition as the number of atoms of each element present in the unit cell. Specifically, for a vocabulary of D elements, where each element can have up to K atoms, we construct D-dimensional vectors where each entry indicates the number of atoms of element d. This yields the space $\mathcal{S}_C = \{(k_1, \dots k_D) | k_d \in \{0, 1, \dots K\}, d = 1, \dots D\}$. The action space of the composition subspace consists of the choice of element and number of atoms, that is $\mathbb{A}_C = \{1, \dots D\} \times \{1, \dots K\}$, plus the special "stop" action to finish a trajectory.

As per the law of electroneutrality, the sum of positive and negative charge in a solid must be equal. Since we have control over the action space of the GFlowNet, we incorporate a hard constraint to ensure that all the generated compositions can have a neutral charge. We argue that this is an advantage over methods that aim at learning such properties implicitly from the data [22].

Finally, we also have an opportunity to incorporate inter-subspace constraints between the space group and the composition. Because of the symmetry imposed by each space group, not all compositions are possible given a space group, and given a composition not all space groups are valid. By way of illustration, the most restrictive space group—international number 230–cannot accommodate compositions with fewer than 16 atoms per element. Since here the trajectories of a Crystal-GFN first sample a space group, we restrict the number of atoms per element in the composition given the space group.

3.3 Lattice parameters

The lattice parameters $(a, b, c, \alpha, \beta, \gamma)$, in combination with the lattice system (or crystal system) determine the shape of the unit cell in three dimensions. While both the composition and the space group subspaces are discrete, the lattice parameters are real-valued and thus yield a continuous sample space $\mathcal{X}_{LP} = \{(\ell_{min}, \ell_{max})^3 \times (\theta_{min}, \theta_{max})^3\}$, where ℓ_{min} and ℓ_{max} are the minimum and maximum lengths and θ_{min} and θ_{max} are the minimum and maximum angles, respectively. In order to sample sets of continuous lattice parameters, the GFlowNet policy model outputs the parameters of a six-dimensional mixture of Beta distributions.

Here, we incorporate an inter-subspace constraint from the space group. Every space group is classified into a lattice system which, in turn, constrains the values of the lattice parameters. For instance, for the cubic lattice system, $a = b = c$ and $\alpha = \beta = \gamma = 90°$. The remaining constraints are illustrated in Fig. 3. This drastically reduces the state space and removes the need of having to learn such relationships and constraints from the data.

3.4 Reward function

As explained in Section 2, our Crystal-GFN is trained to sample proportionally to a reward function. This offers the flexibility of using as reward any quantity of interest, such as a desired property of materials. As an initial proof of concept for the method, we choose the reward to be based on the formation energy (FE) of the sampled crystal structure, in order to generate materials with higher likelihood of being thermodynamically stable. Because the sampled crystal structure may be unknown or not characterised in existing data bases, its true FE may be unknown too. This is why we train a *proxy* machine learning model to predict the FE given a crystal $x \in \mathcal{X}$, which is parameterised as the output of Crystal-GFN. In particular, we train a physics-informed Multi-Layer Perceptron (MLP) on the MatBench [6] data set. Its architecture is detailed in Appendix D.1.

4 Empirical evaluation

In this section, we present the results of experiments designed to serve as proof of concept for Crystal-GFN. The motivation of using GFlowNet to generate crystal structures is two-fold: One, we are interested in discovering structures with high scores of a property of interest, in this case low formation energy. At the same time, we want to discover not just one but multiple and diverse structures. The main reason to seek diverse samples is that often the target function is underspecified and the true objective is multifaceted or unknown. A natural way of dealing with underspecification is to try multiple candidates to increase the likelihood of finding successful structures for the downstream applications [12, 9]. Thus, we here aim at analysing whether Crystal-GFN is able to discover diverse crystal structures with low formation energy.

4.1 Experimental setup

Exploring the infinitely large space of crystals to find the ones with certain properties is a daunting "needle-in-a-haystack" task. In order to make the haystack a little smaller—though still infinitely large—we restrict the search task to a subset of space groups and compositions, and to a limited range of the lattice parameters, as described in Appendix C. We train the Crystal-GFN for 50,000 iterations, which amounts to 500,000 queries to the proxy model and about 12 hours on a CPU-only machine. Further details about the experimental setup are provided in Appendix C.

4.2 Results

In this section, we present results related to our proxy model, to the distribution of the formation energy in Crystal-GFN samples and to the diversity of these samples.

Proxy model of the formation energy As presented in Section 3.4, our reward function is based on a proxy MLP trained on MatBench. Since part of our analysis of the performance of the proposed Crystal-GFN depends on the prediction of the formation energy by this proxy model, it is important to first verify its accuracy. In contrast to other methods, it does not rely on atomic positions to predict the FE of a crystal, but rather on a higher-level description of the crystal, namely its space group, composition and lattice parameters, akin to the outputs of Crystal-GFN. It achieves a mean absolute error (MAE) of 0.10 ± 0.005 eV on the validation set. We further describe its hyper-parameters and detailed performance in Appendices D.2 and D.3.

Predicted formation energy of samples Evaluating generative models is well-known to be a hard, task-dependent problem. Here, in order to gain insights about the sampling policy of the trained Crystal-GFN, we sample 10,000 crystals and compare the distribution with 1) the validation set from MatBench and 2) 10,000 samples from a randomly initialised and untrained Crystal-GFN. Section 4.2 shows the kernel density estimation of the formation energies predicted by our proxy model in the aforementioned sets of samples. As main conclusion, we observe that training the Crystal-GFN shifts the formation energy of the sampled crystals towards lower values, with a median of -3.1 eV, even lower than the median energy in the validation set. Given the vast search space, it is remarkable that with a relatively short training, Crystal-GFN learns to sample from a distribution where 95 % of the structures have a formation energy lower than -2 eV (as predicted by the proxy model).

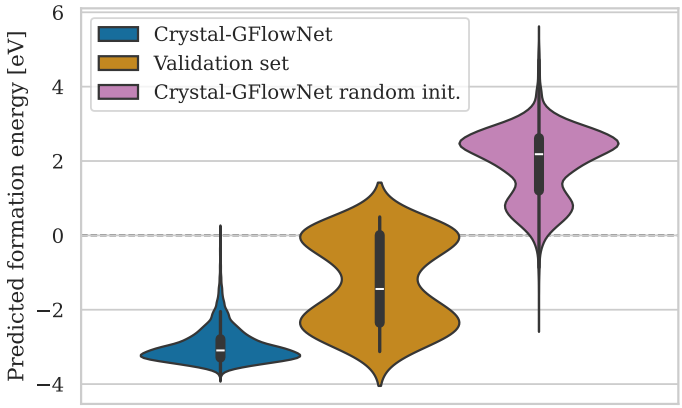


Figure 2: A comparison of the distributions of the formation energy predicted by our proxy model in three relevant distributions of samples: in blue, samples from Crystal-GFN after training; in orange, the validation set, representative of the MatBench database; in pink, samples from an untrained Crystal-GFN. As a main conclusion, we observe that Crystal-GFN, after training, manages to sample crystals with predicted formation energies in the range of the validation set.

Diversity of samples As discussed before, an important goal in certain materials discovery applications is to find diverse candidates. In other words, sampling crystals with low formation energy would be almost useless for most applications if all the crystals were the same or very similar. To gain insights about the diversity, we analyse the crystal structures sampled by the trained Crystal-GFN. The main conclusion is that Crystal-GFN samples candidates with very high diversity. In particular, regarding the composition, we find that all 12 elements are found in the samples from the Crystal-GFN, and 10 of them appear in the 100 samples with lowest predicted energy. Regarding crystal categories and space groups, we find that 5 out of the 8 crystal-lattice systems and all 5 point symmetries are found in just the top-100 samples. While not all 113 space groups are found in the 10,000 samples, we find 80 (70 %) of them. Finally, in terms of lattice parameters, we also observe relatively similar distributions of lengths and angles, compared to the MatBench data set. Further details and visualisations of these results are provided in Appendix E.

5 Conclusions and future work

In this paper, we have introduced a new generative model, Crystal-GFN, to sample inorganic crystal structures proportionally to a property of interest. A key feature of Crystal-GFN is its flexibility to be trained with any available reward function and to incorporate domain constraints. The latter is due to the fact that Crystal-GFN constructs crystals sequentially in the space of space groups, compositions and lattice parameters of the unit cell. This parameterization, inspired by theoretical crystallography, has allowed to flexibly incorporate constraints regarding the neutral charge of the composition, the compatibility between composition and space group and between space group and lattice parameters.

By training a Crystal-GFN with a reward function based on the formation energy predicted by a proxy model trained on the MatBench database, we have shown it learns to sample candidates with low predicted formation energy (lower mean and median that the distribution of the MatBench data set). Importantly, it does so by sampling highly diverse crystals, in terms of coverage of space groups, compositions and lattice parameters.

Interesting directions for future work include extending Crystal-GFN with more domain-inspired constraints as well as with additional subspaces to sample the atom positions in the unit cell. Alternatively, crystal structure prediction from the samples could be performed using pre-trained machine learning models that generate atomic coordinates. Furthermore, it would be interesting to explore other properties of interest beyond the formation energy of the sampled crystals.

Acknowledgments and Disclosure of Funding

We thank Yasmine Benabed for her contribution to the conception and first steps of this project. We also thank Bruno Rousseau, Simon Blackburn and Mickael Dollé for their valuable insights from the materials science perspective. Sékou-Oumar Kaba and Michael Kilgour have also participated in helpful discussions. We thank Mila's IDT team for their support. Finally, we thank Quebec's Ministry of Economy, Innovation and Energy for their financial support.

This research was enabled in part by compute resources provided by Mila (mila.quebec).

References

[1] Rasool Ahmad and Wei Cai. Free energy calculation of crystalline solids using normalizing flows. *Modelling and Simulation in Materials Science and Engineering*, 30(6):065007, 2022.

[2] Emmanuel Bengio, Moksh Jain, Maksym Korablyov, Doina Precup, and Yoshua Bengio. Flow network based generative models for non-iterative diverse candidate generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

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

[4] Guanjian Cheng, Xin-Gao Gong, and Wan-Jian Yin. Crystal structure prediction by combining graph network and optimization algorithm. *Nature Communications*, 13(1), March 2022.

[5] Callum J Court, Batuhan Yildirim, Apoorv Jain, and Jacqueline M Cole. 3-d inorganic crystal structure generation and property prediction via representation learning. *Journal of Chemical Information and Modeling*, 60(10):4518–4535, 2020.

[6] Dunn, A., Wang, Q., Ganose, A., Dopp, D., Jain, and A. Benchmarking materials property prediction methods: The matbench test set and automatminer reference algorithm., 2020. Accessed 2023-09-28.

[7] Alexandre Duval, Victor Schmidt, Santiago Miret, Yoshua Bengio, Alex Hernández-García, and David Rolnick. Phast: Physics-aware, scalable, and task-specific gnns for accelerated catalyst design. *arXiv preprint arXiv: 2211.12020*, 2022.

[8] Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Advances in neural information processing systems*, 32, 2019.

[9] Alex Hernandez-Garcia, Nikita Saxena, Moksh Jain, Cheng-Hao Liu, and Yoshua Bengio. Multi-fidelity active learning with gflownets. *arXiv preprint arXiv: 2306.11715*, 2023.

[10] Jordan Hoffmann, Louis Maestrati, Yoshihide Sawada, Jian Tang, Jean Michel Sellier, and Yoshua Bengio. Data-driven approach to encoding and decoding 3-d crystal structures. *arXiv preprint arXiv:1909.00949*, 2019.

[11] Moksh Jain, Emmanuel Bengio, Alex Hernandez-Garcia, Jarrid Rector-Brooks, Bonaventure FP Dossou, Chanakya Ajit Ekbote, Jie Fu, Tianyu Zhang, Michael Kilgour, Dinghuai Zhang, et al. Biological sequence design with GFlowNets. In *International Conference on Machine Learning (ICML)*, volume 162. PMLR, 2022.

[12] Moksh Jain, Tristan Deleu, Jason Hartford, Cheng-Hao Liu, Alex Hernandez-Garcia, and Yoshua Bengio. GFlowNets for AI-driven scientific discovery. *Digital Discovery*, 2023.

[13] Sungwon Kim, Juhwan Noh, Geun Ho Gu, Alan Aspuru-Guzik, and Yousung Jung. Generative adversarial networks for crystal structure prediction. *ACS central science*, 6(8):1412–1420, 2020.

[14] Teng Long, Nuno M Fortunato, Ingo Opahle, Yixuan Zhang, Ilias Samathrakis, Chen Shen, Oliver Gutfleisch, and Hongbin Zhang. Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Computational Materials*, 7(1):66, 2021.

[15] Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in GFlowNets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

[16] Asma Nouira, J. Crivello, and Nataliya Sokolovska. Crystalgan: Learning to discover crystallographic structures with generative adversarial networks. *AAAI Spring Symposium Combining Machine Learning with Knowledge Engineering*, 2018.

[17] Teerachote Pakornchote, Natthaphon Choomphon-anomakhun, Sorrjit Arrerut, Chayanon Atthapak, Sakarn Khamkaeo, Thiparat Chotibut, and Thiti Bovornratanaraks. Diffusion probabilistic models enhance variational autoencoder for crystal structure generative modeling. *arXiv preprint arXiv:2308.02165*, 2023.

[18] Zekun Ren, Juhwan Noh, Siyu Tian, Felipe Oviedo, Guangzong Xing, Qiaohao Liang, Armin Aberle, Yi Liu, Qianxiao Li, Senthilnath Jayavelu, et al. Inverse design of crystals using generalized invertible crystallographic representation. *arXiv preprint arXiv:2005.07609*, 3(6):7, 2020.

[19] Zekun Ren, Siyu Isaac Parker Tian, Juhwan Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G Aberle, Shijing Sun, et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 5(1):314–335, 2022.

[20] Victor Garcia Satorras, E. Hoogeboom, F. Fuchs, I. Posner, and M. Welling. E(n) equivariant normalizing flows. *Neural Information Processing Systems*, 2021.

[21] 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.

[22] Tian Xie, Xiang Fu, O. Ganea, R. Barzilay, and T. Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *International Conference On Learning Representations*, 2021.

[23] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

[24] Yong Zhao, Mohammed Al-Fahdi, Ming Hu, Edirisuriya MD Siriwardane, Yuqi Song, Alireza Nasiri, and Jianjun Hu. High-throughput discovery of novel cubic crystal materials using deep generative neural networks. *Advanced Science*, 8(20):2100566, 2021.

[25] Shuxin Zheng, Jiyan He, Chang Liu, Yu Shi, Ziheng Lu, Weitao Feng, Fusong Ju, Jiaxi Wang, Jianwei Zhu, Yaosen Min, et al. Towards predicting equilibrium distributions for molecular systems with deep learning. *arXiv preprint arXiv:2306.05445*, 2023.

A Related Work

Material and Molecule generation. There exists a myriad of generative methods for molecular and crystal generation, generally falling into one of these catagories: Variational Auto Encoders (VAE) [19, 17], Generative Adversarial Networks (GANs) [3, 16, 14], Normalizing flows [20, 1] and Diffusion models [25, 23]. In general, these methods are trained to reconstruct the training data from a latent distribution by maximising its likelihood or minimising the discrepancy between generated sampled and the real distribution. Once trained, they can generate new instances by sampling from the learned latent space.

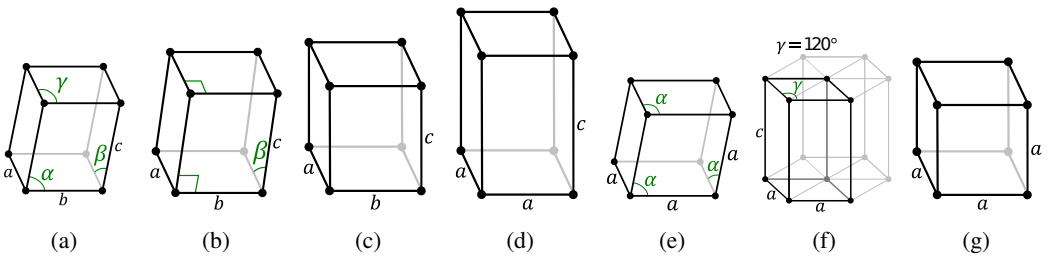
[10, 14, 5] treat materials as 3D voxel images, but the process of decoding images back to atom types and coordinates often results in low validity, and the models are not rotationally invariant. Several methods [18, 13, 24] encode unit cell related information to generate crystals. GANCSP [13] and CubicGAN [24] use as input fractional coordinates, element properties, and lattice parameters, to build models that generate crystals conditioned on composition or both composition and space group. Such models are generally not invariant to any Euclidean transformations. One close baseline, CDVAE [22], employs a diffusion model in conjunction with a VAE to sample new materials, leveraging lattice parameters. However, it is does not enable one to impose domain-expertise based constraints into the generation process. [4] leveraged physics and symmetry-based constraints to solve an optimization problem over crystal graphs to obtain atomic coordinates, but requires a composition and lattice information as input.

Finally, auto-regressive methods like G-SchNet [8] or G-SphereNet generates 3D molecules sequentially, placing atom-by-atom. They are unaware of periodicity and cannot generate the lattice, making them ill-suited for crystals. This also applies to flow models [21].

As these models directly sample 3D positions, they struggle to embed essential physical constraints[3], often leading to physically unrealistic and unstable structures. Unlike them, by selecting the composition, lattice, and space group and imposing validity constraints, our approach inherently incorporates domain knowledge about materials science and crystallography ensuring that the generated crystals adhere to fundamental principles and constraints, resulting in more physically valid structures.

GFlowNets: GFlowNets are a new [2] class of trained generative models that can be used to sample objects through a series of constructive steps, allowing for high reward samples and also diversity in sampling. They have the potential of being employed in a variety of different materials discovery avenues [12], starting from sample design to experiment design. In particular, they have already shown promise in the area of drug-discovery, having been used to sequentially construct small molecules[2] and biological sequences such as peptides and DNA/protein sequences[11].

B Crystals



(a) (b) (c) (d) (e) (f) (g)

Figure 3: The seven lattice systems and the constraints they impose on the lattice parameters. In order a–g: triclinic, monoclinic, orthorhombic, tetragonal, rhombohedral, hexagonal and cubic. Source of the figures: Wikimedia Commons, licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license.

Materials encompass the substances that make up objects. Their distinctive characteristics arise from a combination of factors, including their composition, structure, and manufacturing processes.

[3] like energy minimization, dihedral angles geometric constraints or bonding preferences between atom types.

While materials come in various forms, our primary focus will be on solid-state materials, specifically crystals, owing to their significance in many machine learning applications we have encountered.

Solid state materials consist of a multitude of smaller building blocks, such as atoms or molecules, securely positioned in fixed locations. They exhibit high-density packing and strong mutual attraction, resulting in a stable structure with a well-defined volume. Materials that possess a well-ordered, infinitely repetitive structure are categorized as crystals, while those lacking such long-range positional order are termed amorphous.

Crystals, also known as crystalline materials, represent a category of solid substances characterised by their periodic structure. This entails a consistent and predictable repetition of atom or molecule placement within the crystal across all spatial dimensions. It is this inherent periodicity that sets crystals apart from smaller molecules and confers unique properties upon crystals. They manifest in diverse forms and showcase exceptional attributes like transparency and elevated melting points. The study of crystals offers a pathway to precise characterization and controlled manipulation, opening up avenues for engineering materials with targeted properties to fulfill a wide range of needs. The subsequent section delves into how machine learning models address and harness this periodicity.

A crystal structure includes the type and position of every atom as well as the translational axes which allow the structure to repeat. Since crystal structures extend infinitely, we need to define some small section that can be repeated. We call this small section a *unit cell*. Unit cells are made of a lattice and basis. Lattices tell you how the crystal is repeated, the basis tells you what is repeated. There are six parameters of a unit cell: 3 edges a, b, c and the angles between the edges α, β, γ. The edges of a unit cell may be or may not be perpendicular to each other. A crystal structure combines the *lattice* (how atoms are translated) with a *basis* (which atoms are translated) which theoretically describes every atom in the crystal. The lattice is just a mathematical idea, but the basis is the "real" component which is repeated according to the mathematical lattice.

Lattice The lattice is fully specified by its basis vectors $\mathbf{a}_i$ and is associated with a group of translations T_Λ. The lattice Λ is the underlying structure of the *unit cells*, which have the property of tilling the space $\mathbb{R}^n$ when translated by lattice vectors. Of particular prominence is the primitive cell U for the basis $\mathbf{a}_i$: $U \doteq \{\sum_i^n x_i \mathbf{a}_i \mid 0 \leq x_i < 1\}$.

In a material, the unit cell comprises a set of atomic positions $S = \{(Z_i, \mathbf{x}_i) \mid \mathbf{x}_i \in U\}$, where $\mathbf{x}_i$ is the atom position (called the basis vector) and Z_i its atomic number. S can contain an arbitrary number of atoms and is not constrained to a particular structure.

Periodic boundary conditions (pbc) are commonly used to represent the infinite repeating nature of crystals and other periodic systems. Rather than confining the analysis to a single unit cell housing all atoms, pbc envision an infinite array of identical unit cells extending indefinitely in all directions. This means that if an atom or molecule exits the cell on one side, it effectively re-enters from the opposite side, as if traversing a continuous lattice. This approach enables scientists to simulate the behavior of an infinitely extending crystal using a finite computational domain.

C Experimental setup

In this section, we provide additional details about the hyper-parameters and experimental setup used to train the Crystal-GFN to obtain the results presented in Section 4.

In order to reduce the search space in our experiments, we apply the following restrictions:

- Compositions consist of up to 5 different elements from the subset of these 12 elements: H, Li, C, N, O, F, Mg, Si, P, S, Cl and Fe. These are the most common elements in the MatBench data set used to train the proxy model.
- Compositions can contain up to 50 atoms in total and up to 16 atoms per element. This number is obtained by finding the lowest Wyckoff position multiplicity for each space group, and then computing the maximum of these values across all space groups. This means that 16 is the lowest possible value that still makes it possible for the Crystal-GFN to generate samples from all space groups while respecting their symmetry constraints.
- Space groups are the intersection of train and validation space groups, from the proxy model's data set. There are 113 of them.

- The minimum and maximum lengths of the unit cell are 0.9 and 100 angstroms, respectively; the minimum and maximum angles are 50° and 150°, respectively.

We train the Crystal-GFN by sampling 10 trajectories per iteration from the current forward policy. In order to encourage further exploration, 10 % of the steps in the trajectories are sampled at random from a uniform distribution. In total, we train for 50,000 iterations, which amounts to 500,000 queries to the proxy model. This took about 12 hours on a CPU-only machine.

The architecture of both the forward and the backward GFlowNet policy models is a 3-layer multi-layer perceptron with 256 units per layer. We trained with the Adam optimiser and a learning rate of 0.0001. As is common with the Trajectory Balance objective [15], we set a higher learning rate (0.01) for the 16 learning weights used to parameterised partition function. In order to sample structures with lower (negative) formation energy, we set a temperature of 8 in the reward function defined in Section 3.4.

The distribution to sample the increments of the lattice parameters subspace is a mixture of 5 Beta distributions. In order to ensure numerical stability during training, we restrict the values of the coefficients of the Beta distributions to the range $[0.1, 100]$. One of the conditions that must be satisfied by generalised GFlowNets is that trajectories must have finite length. To this end we set a minimum increment of 10 % the range of each dimension.

D Proxy MLP

D.1 Architecture

For a given crystal structure, the input to the proxy MLP is the concatenation of: a physical embedding of the crystal's elements using PhAST [7], a learned embedding for its space group and the standardized lattice parameters. We stratify the MatBench data set into train, validation and test sets, controlling for the distribution in target FE. The final reward function is a Boltzmann version of the proxy: $R(x) = \exp\big(-\frac{\text{MLP}(x)}{T}\big)$ where T is a temperature hyper-parameter. This ensures that lower FE yields higher positive reward, and we can control our preference for lower energies with the temperature T.

Overall, the architecture consists of concatenated representations for the composition (h_C), the space group h_{SG} and the lattice parameters (h_{LP}) that are used as input to a prediction MLP $\hat{y} = \text{MLP}_{out}([h_C, h_{SG}, h_{LP}])$. In the following we describe how each h is obtained from the data:

- $\mathbf{h_C}$: Each element Z_i in the composition is embedded with PhAST [7]: it is a concatenation of 1/ static physical properties P_Z projected into some latent space with a small MLP 2/ a learned embedding for the atomic number Z, Period P and the Group G of the element, noted E_i:

$$h_C = \text{MLP}_C\Big(\big[\text{MLP}_P(P_Z), E_Z, E_P, E_G\big]\Big) \tag{1}$$

- $\mathbf{h_{SG}}$: Each space group is embedded as a lookup in a learnable embedding matrix.
- $\mathbf{h_{LP}}$: Lattice parameters are first standardize element-wise using the training set statistics. This 6-dimensional vector is then passed through a small MLP:

$$h_{LP} = \text{MLP}_{LP}\big(\frac{\text{LP}(x) - \mu_{train}(\text{LP})}{\sigma_{train}(\text{LP})}\big) \tag{2}$$

D.2 Performance

Our proxy MLP model achieves an overall Mean Absolute Error of 0.10 ± 0.005 eV[4]. For reference, best performing methods in Matbench leaderboard[5] can achieve up to $10\times$ better performance.

[4] 95% confidence interval, computed modeling the MAE as a log-normal distribution. This is further justified by Fig. 5.

[5] `https://matbench.materialsproject.org/Leaderboards%20Per-Task/matbench_v0.1_matbench_mp_e_form/`

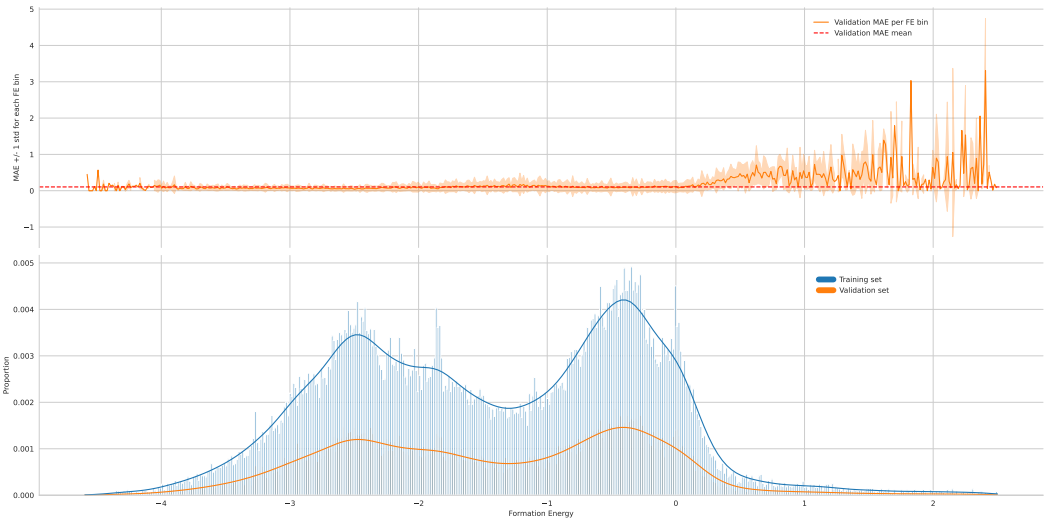


Figure 4: Proxy MLP performance on the validation set and data split FE distributions. The average MAE is 0.10eV. We can also see the effect of the stratification algorithm wich yields similar FE distributions between the train and validation data set splits.

However, they all use 3D positions of the crystals, and remarkably, our approach outperforms some of them, indicating that leveraging composition, space-group and lattice parametrs serves as a viable representation in the formation energy estimation task.

In Figures 4 and 5 we detail further the performance of our Proxy MLP on the MatBench validation set. In particular we observe that it maintains good performance for structures with FE up to $\sim 0.3ev$. The MAE beyond that FE level becomes high and unstable, but this is expected as those are the regions of the target space with least data available. Additionally we can verify in Figure 4 that our stratification algorithm works as expected, with similar FE distributions in the validation and train splits.

D.3 Hyper parameters

We present the hyper parameters of our Proxy MLP in Table 1 along with a description of their role in the architecture.

E Additional results

In this section, we provide additional results demonstrating the diversity of the samples generated by the Crystal-GFN. They are displayed in Figure 6, Figure 8, Figure 9 and Figure 10.

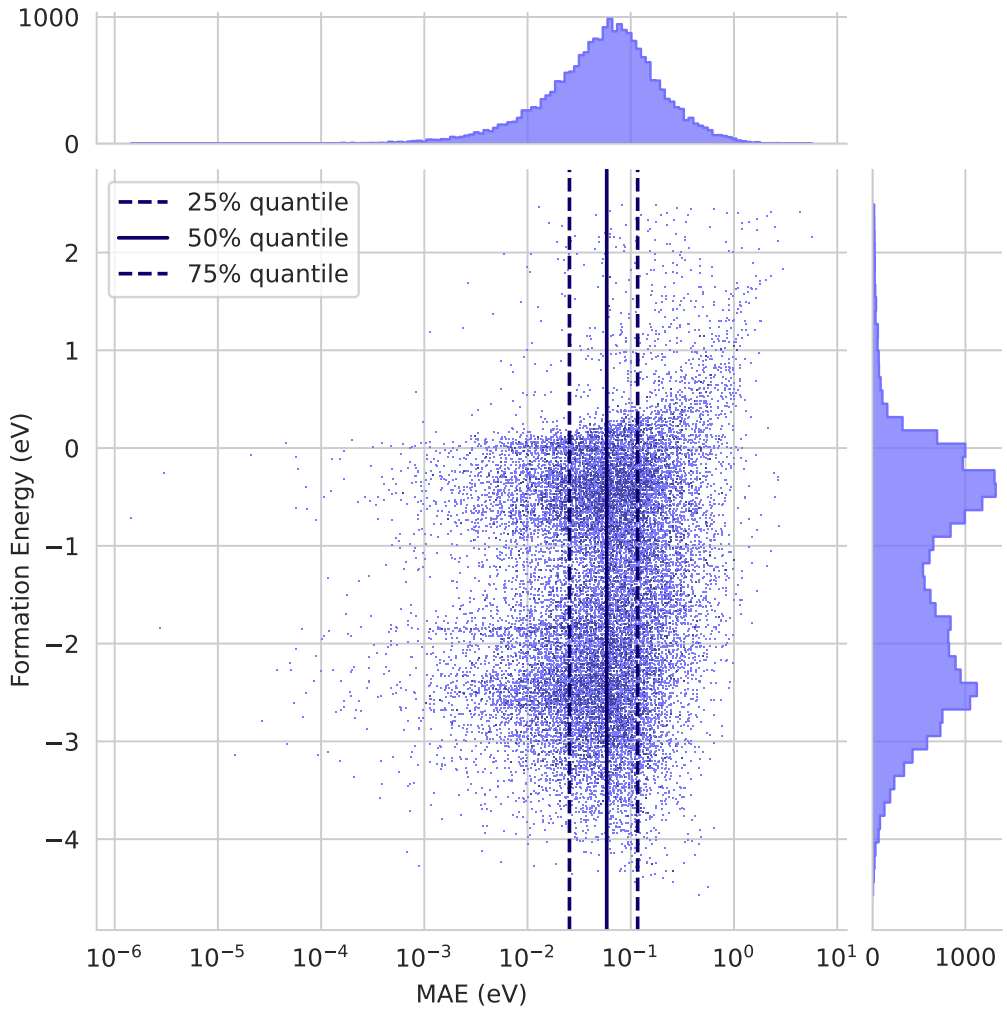


Figure 5: Distribution of FE values in the validation set and associated Proxy MLP MAE, with $25\%, 50\%$ and 75% MAE quantiles.

Hyper parameter	**Value**	**Description**
properties_proj_size	64	Projection size of atomic properties $h = \mathbf{W}p_Z$
group_emb_size	16	Embedding size of element's group
period_emb_size	256	Embedding size of element's period
z_emb_size	128	Embedding size of element's atomic number
sg_emb_size	128	Crystal space group embedding size
lat_hidden_channels	284	Hidden channels for layers processing the lattice parameters
lat_num_layers	1	Number of hidden layers for the lattice parameters MLP
num_layers	5	Number of layers for final MLP processing composition, space group and lattice parameters hidden representations
hidden_channels	576	Size of final MLP hidden layers.
optimizer	Adam	
lr	0.0017	Learning rate
batch_size	448	
scheduler	ReduceLROnPlateau	Divide learning rate by 2 when validation MAE does not improve for 4 epochs
es_patience	11	Early stopping patience (in epochs)
epochs	100	Total epochs (effective after early stopping: 97)

Table 1: Proxy MLP hyper parameters

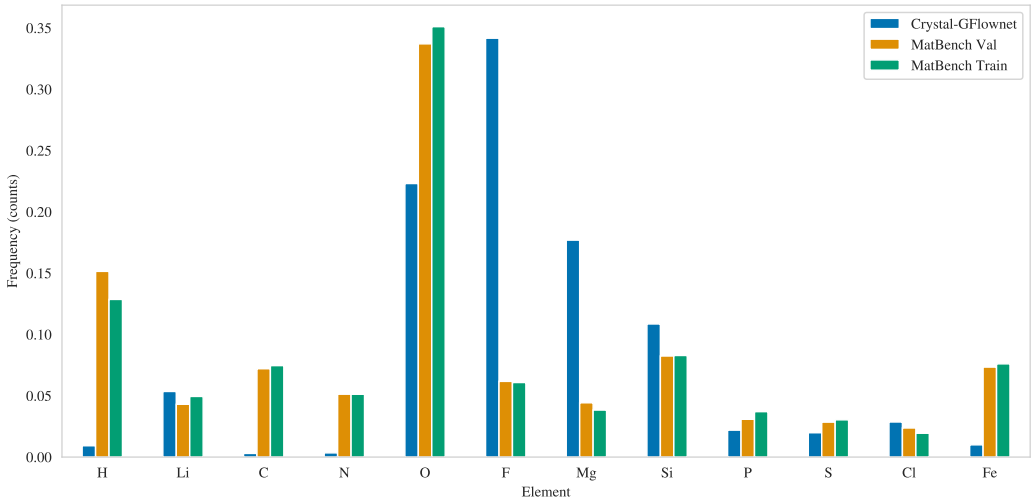


Figure 6: Distribution of total element prevalence per data set (10k Crystal-GFN samples, MatBench val, MatBench train). Each element is counted as per its stoichiometry in the crystal.

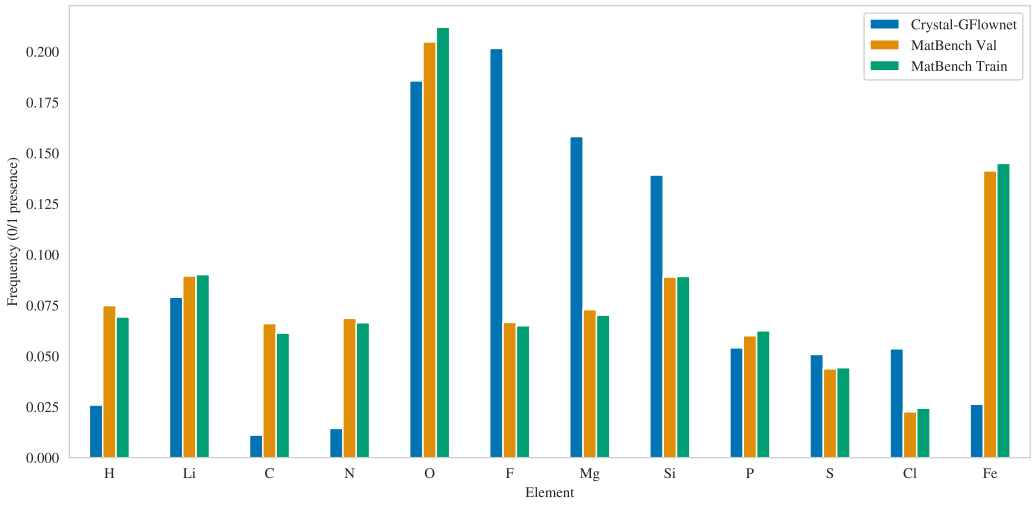


Figure 7: Distribution of binary element prevalence per data set (10k Crystal-GFN samples, MatBench val, MatBench train). Each element is counted only once per crystal.

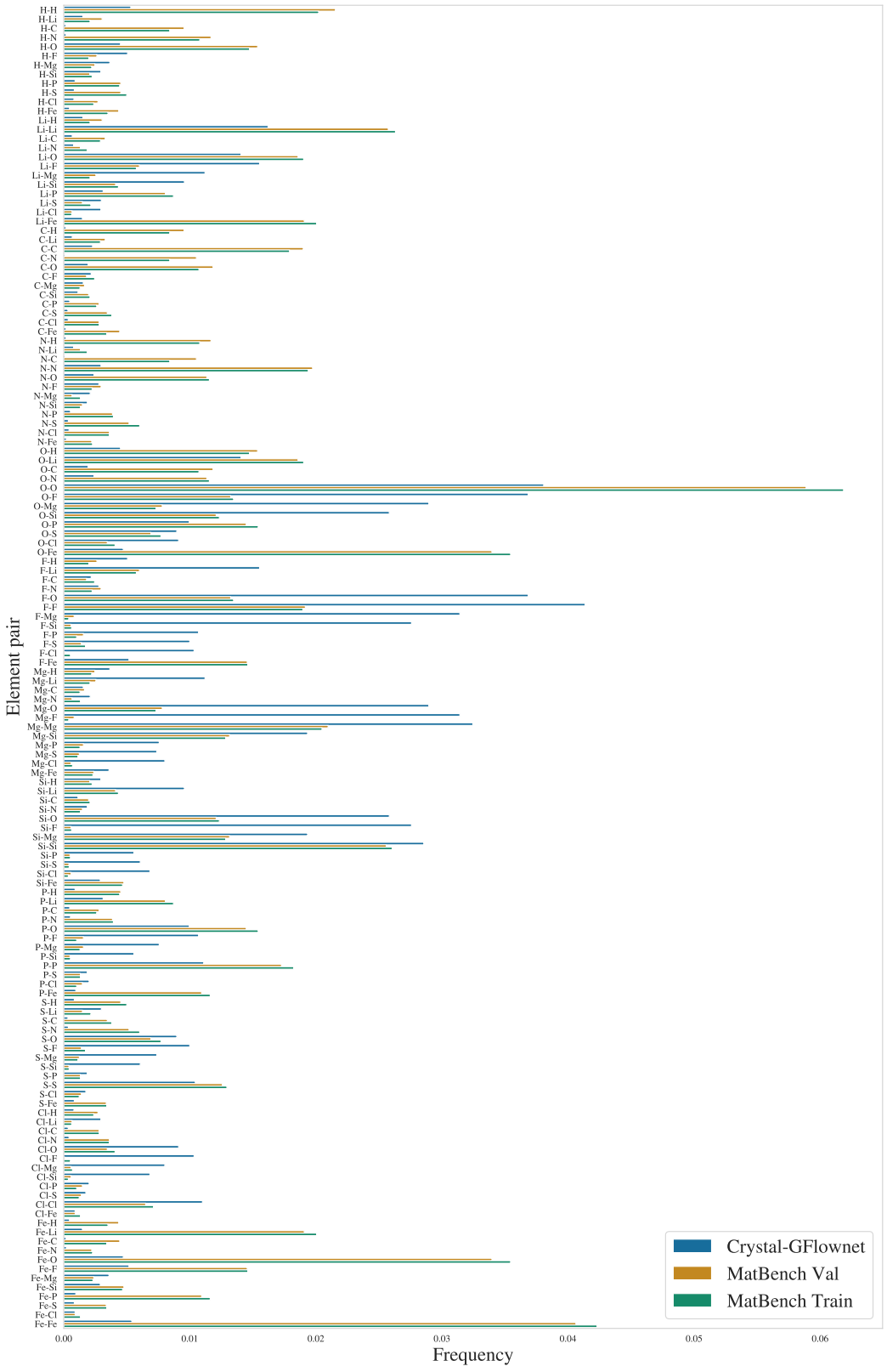


Figure 8: Distribution of binarized element co-occurrences per data set (10k Crystal-GFN samples, MatBench val, MatBench train). A binarized co-occurrence is defined as element i and element j both being present in a crystal, regardless of their respective stoichiometry in that crystal.

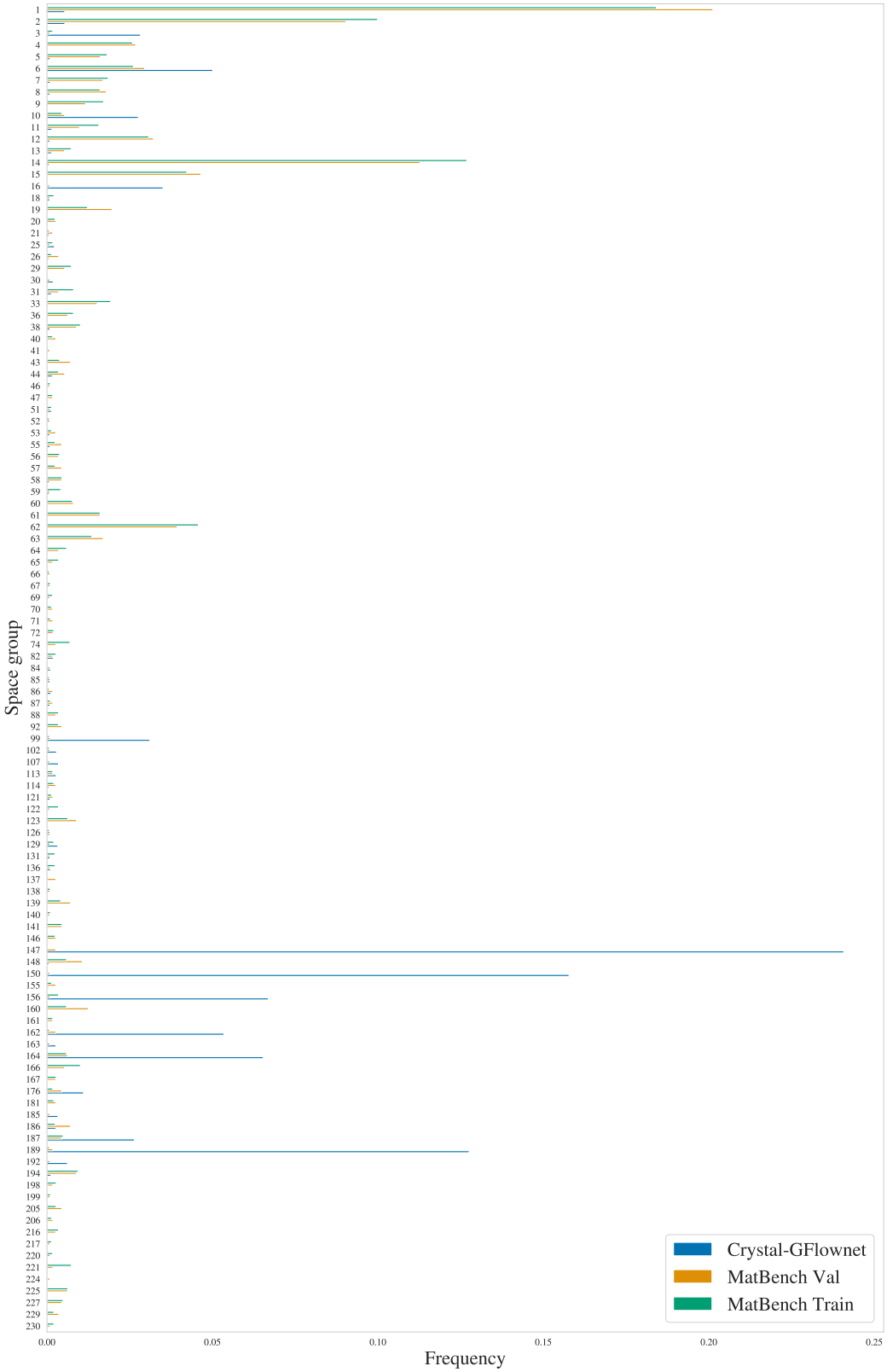


Figure 9: Distribution of Space Groups per data set (10k Crystal-GFN samples, MatBench val, MatBench train)

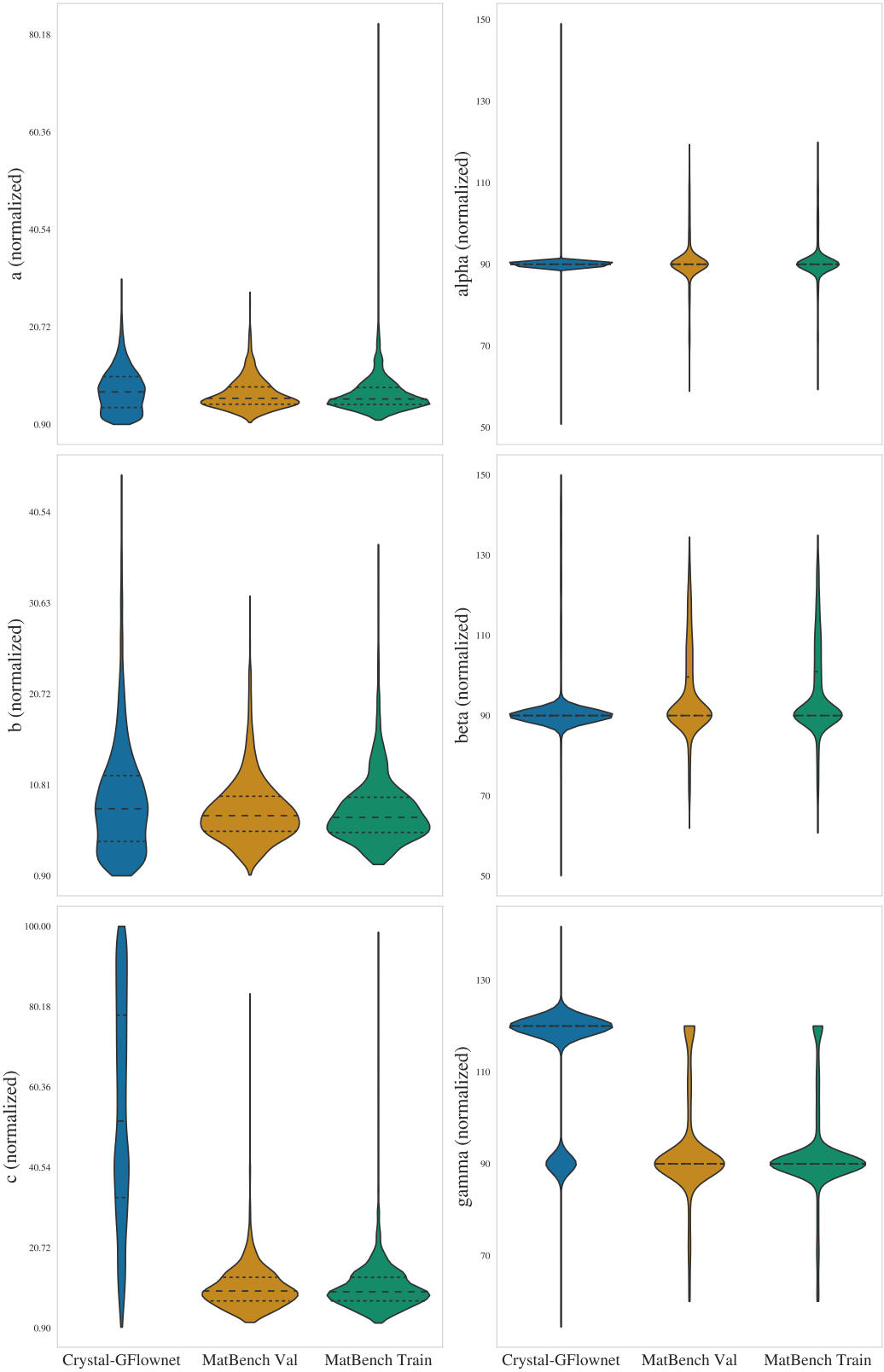


Figure 10: Distribution of lattice parameters