1 Supplementary Information

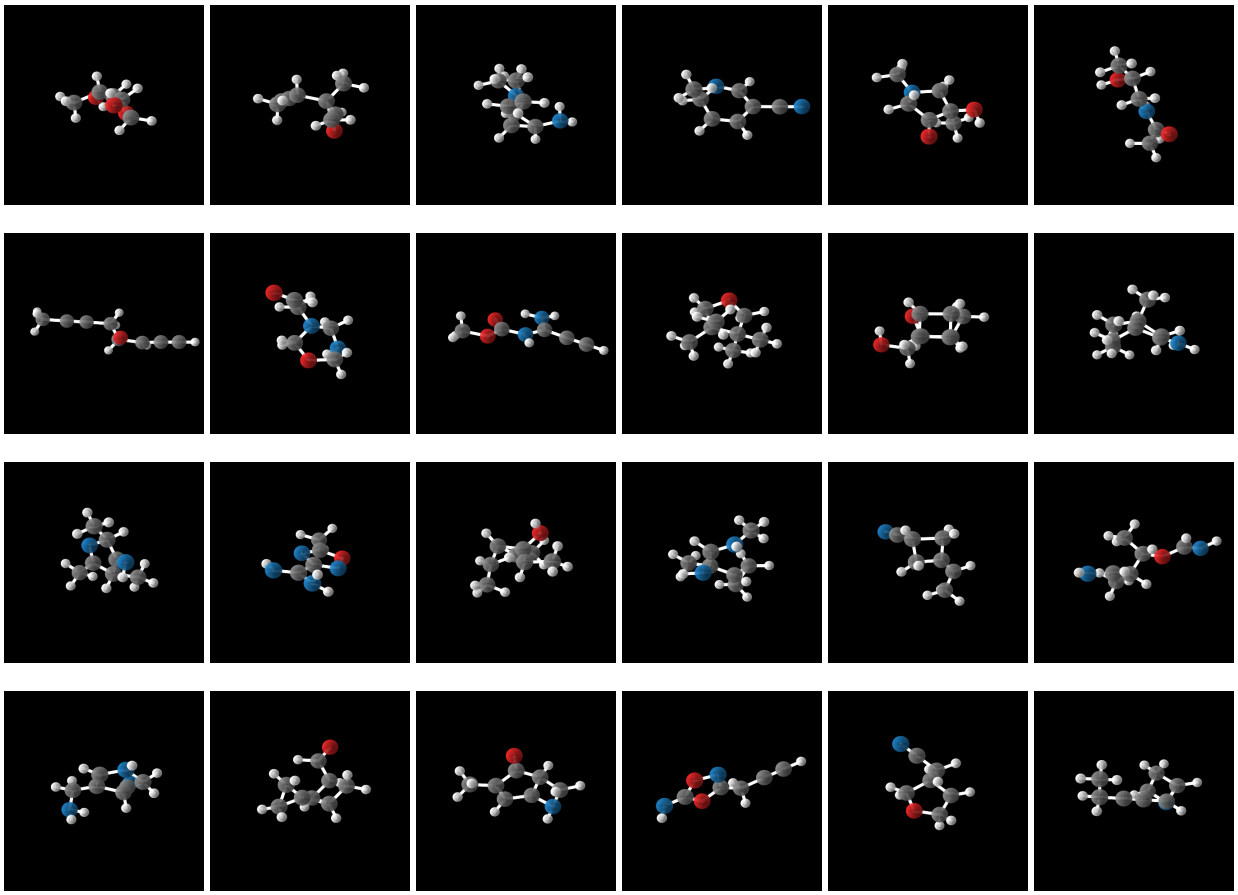


Figure S1. Randomly sampled molecules generated by MolCode trained on QM9.

Table S1. Substructure analysis of the generated molecules. The KL divergence of the bond lengths (upper part) and bond angles (lower part) between the training set and the generated molecules are shown below.

Distances/Angles	E-NFs[60]	G-SchNet[24]	G-SphereNet[43]	EDM[42]	MolCode (w/o bond)	MolCode
CC	0.53	0.44	0.30	0.36	0.32	**0.24**
CN	0.87	0.68	0.45	0.37	0.43	**0.30**
CO	0.49	0.32	0.24	0.26	0.25	**0.21**
NO	0.39	0.27	0.20	0.24	0.19	**0.17**
CCC	1.25	0.96	0.65	0.48	0.66	**0.25**
CCO	0.98	0.85	0.41	0.33	0.47	**0.23**
CNC	1.44	1.37	0.71	0.56	0.64	**0.42**
CCN	1.30	0.95	0.74	0.84	0.62	**0.37**

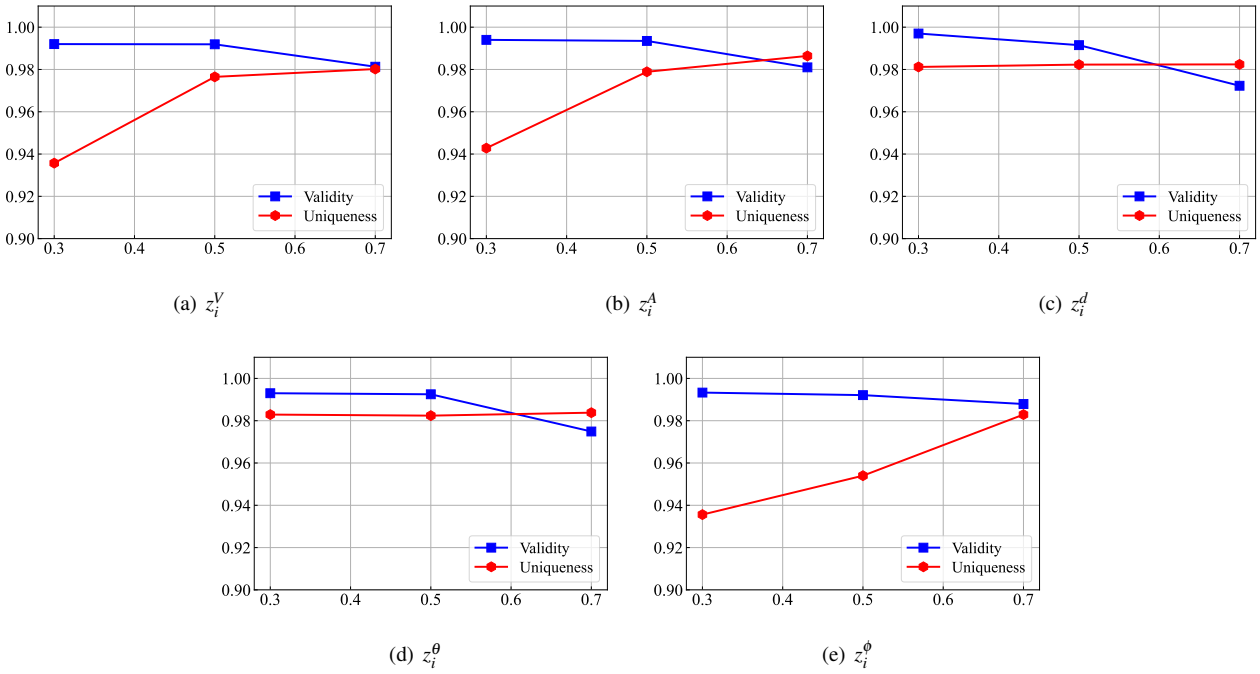
Table S2. Properties of the test set molecules and the generated molecules by different methods in structure-based drug design (the upper part shows the results of baselines and the test set; the lower part shows MolCode and its variants). We report the means and standard deviations. The best results are bolded.

Method	Vina Score (kcal/mol, ↓)	High Affinity(↑)	QED (↑)	SA (↑)	LogP	Lip. (↑)	Sim. (↓)	Div. (↑)
Testset	-7.143±1.76	-	0.477±0.18	0.642±0.17	0.924±2.06	4.443±1.18	-	-
LiGAN[74]	-6.184±0.96	0.146±0.16	0.240±0.17	0.482±0.13	-0.125±2.48	4.146±1.29	0.412±0.17	0.580±0.09
AR[44]	-6.237±1.24	0.188±0.18	0.581±0.13	0.640±0.17	0.232±1.78	4.662±0.54	0.423±0.18	0.733±0.11
GraphBP[46]	-6.368±1.65	0.236±0.15	0.454±0.20	0.459±0.18	0.512±2.13	4.574±0.44	0.413±0.11	0.687±0.15
Pocket2Mol[47]	-7.206±1.82	0.544±0.14	0.649±0.16	0.704±0.18	1.134±1.88	4.905±0.14	0.388±0.21	0.650±0.16
MolCode	**-7.412±1.74**	**0.618±0.15**	**0.652±0.19**	**0.708±0.16**	1.225±2.16	**4.924±0.08**	**0.371±0.20**	**0.693±0.14**
w/o check	-7.076±1.48	0.472±0.13	0.663±0.15	0.664±0.15	1.054±1.85	4.912±0.12	0.382±0.19	0.686±0.12
w/o bond	-6.895±1.67	0.414±0.14	0.620±0.17	0.642±0.13	0.926±1.93	4.829±0.18	0.390±0.21	0.676±0.16

In Table.S2, we show more qualitative evaluations of the generated ligand molecules in structure-based drug design. We choose metrics that are widely used in previous works[44,46,47] to evaluate the qualities of the sampled molecules: (1) **Vina Score** measures the binding affinity between the generated molecules and the protein pockets; (2) **High Affinity** is calculated as the percentage of pockets whose generated molecules have higher affinity to the references in the test set; (3) **QED**[68] measures how likely a molecule is a potential drug candidate; (4) **Synthesizability (SA)** represents the difficulty of drug synthesis (the score is normalized between 0 and 1 and higher values indicate easier synthesis); (5) **LogP** is the octanol-water partition coefficient (LogP values should be between -0.4 and 5.6 to be good drug candidates[82]); (6) **Lipinski (Lip.)** calculates how many rules the molecule obeys the Lipinski's rule of five[83]; (7) **Sim.** represents the Tanimoto similarity[84] with the most similar molecules in the training set; (8) **Diversity (Div.)** measures the diversity of generated molecules for a binding pocket (It is calculated as 1 - average pairwise Tanimoto similarities). In our work, The Vina Score is calculated by QVina[69,70] and the chemical properties are calculated by RDKit[71,72] over the valid molecules. Before feeding to Vina, all the generated molecular structures are firstly refined by universal force fields[73]. To sample molecules, we apply 5 independent runs with random seeds.

Table S3. Results of random molecule generation on GEOM-Drugs. We follow the preprocessing procedure in previous work[76]. Validity calculates the percentage of valid molecules among all the generated molecules; Uniqueness refers to the percentage of unique molecules among the valid molecules; Novelty measures the fraction of molecules not in the training set among all the valid and unique molecules. Time records the sampling time for 10,000 molecules. The best results are bolded.

Method	Validity	Uniqueness	Novelty	Time
G-SphereNet[43]	39.16%	96.45%	98.70%	**2746**
EDM[42]	44.71%	**100.00%**	**100.00%**	25284
MolCode (w/o check)	72.30%	98.75%	97.43%	3162
MolCode (w/o bond)	63.46%	98.14%	98.77%	3488
MolCode	**98.82%**	**100.00%**	**100.00%**	3950



(a) z_i^V (b) z_i^A (c) z_i^d

(d) z_i^θ (e) z_i^ϕ

Figure S2. Influence of temperature hyperparameters on Validity and Uniqueness in the random molecule generation task.

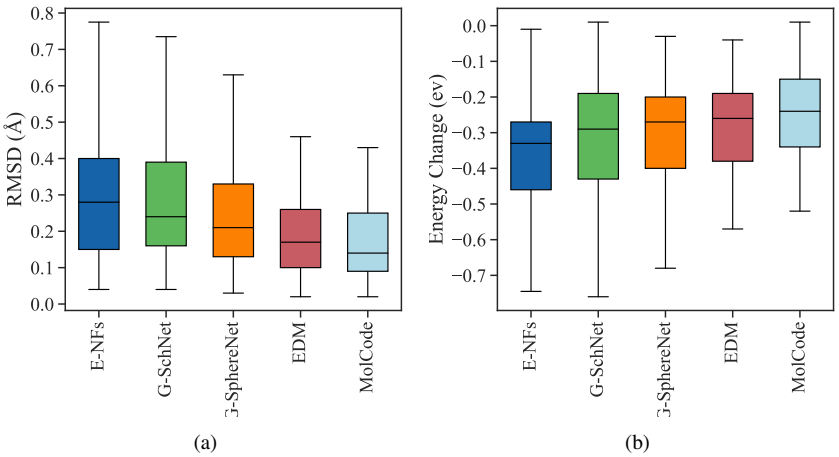


Figure S3. (a) The RMSD of the atomic positions between the generated and the corresponding relaxed molecules with force fields. Following previous works on molecular conformation generation[38,85], we use the empirical force field[86] to optimize molecular conformations. (b) The internal energy change between the generated and the corresponding relaxed molecules ($E_{relaxed} - E_{generated}$). Following cG-SchNet[26] and G-SchNet[24], we use the pretrained SchNet[87] models from SchNetPack[88] to predict the internal energy at zero Kelvin of generated molecules.

Algorithm 1 Training Algorithm of MolCode

Input: Molecular dataset $\mathscr{M}$, learning rate η, Adam hyperparameters β_1, β_2, batch size B, GoGen model with trainable parameter w, latent distribution $p_{Z_V}, p_{Z_A}, p_{Z_d}, p_{Z_\theta}, p_{Z_\phi}$, maximum number of atoms n
Initial: Parameters w of MolCode

while w is not converged **do**
  Sample a batch of B molecule mol from dataset $\mathscr{M}$
  $L = 0$
  for $G \in mol$ **do**
    Set n as the number of atoms in G and order the atoms in G
    for $i = 1, ..., n-1$ **do**
      Get V_i, d_i, θ_i (if $i \geq 2$), ϕ_i (if $i \geq 3$) and the reference atoms (f, c, e)
      Get z_i^V, z_i^d, z_i^θ (if $i \geq 2$), z_i^ϕ (if $i \geq 3$) with the flow modules in MolCode
      $L = L - log p_{Z_V}(z_i^V) - log p_{Z_d}(z_i^d)$
      $L = L - log p_{Z_V}(z_i^\theta)$ (if $i \geq 2$)
      $L = L - log p_{Z_V}(z_i^\phi)$ (if $i \geq 3$)
      for $j \in \{f, c, e\}$ **do**
        Get A_{ij} and z_{ij}^A
        $L = L - log p_{Z_A}(z_{ij}^A)$
      end for
      Add the binary cross entropy loss for the focal atom selection to L
    end for
  end for
  $w \leftarrow \text{ADAM}(\frac{L}{B}, w, \eta, \beta_1, \beta_2)$
end while

Algorithm 2 Generation Algorithm of MolCode

Input: GoGen model with parameter w, latent distribution $p_{Z_V}, p_{Z_A}, p_{Z_d}, p_{Z_\theta}, p_{Z_\phi}$, maximum number of atoms n, maximum number of trials to sample bond types T

for $i = 1, ..., n-1$ **do**
  Initialize molecular graph G_1 with one carbon atom, whose coordinate is R_0 = [0, 0, 0]
  Sample $z_i^V \sim p_{Z_V}$ and generate V_i
  Get the candidate focal atom set by the atom-wise classifier
  Get the reference atoms $\{f, c, e\}$
  for $j \in \{f, c, e\}$ **do**
    Count = 0
    Get $z_{ij}^A \sim p_{Z_A}$ and generate A_{ij}
    if $\sum_j |A_{ij}| \geq$ Valency(X_i) or $\sum_i |A_{ij}| \geq$ Valency(X_j) and Count $\leq T$ **then**
      Reject A_{ij} and sample a new z_{ij}^A; Count+=1
    else
      Assign no bond to A_{ij}
    end if
  end for
  if the candidate focal atom set is empty or $\sum_j |A_{ij}| = 0$ **then**
    Output G_i
  else
    Random select the focal atom f from the candidate focal atom set
  end if
  Sample z_i^d, z_i^θ (if $i \geq 2$), z_i^ϕ (if $i \geq 3$)
  Generate d_i, θ_i (if $i \geq 2$), ϕ_i (if $i \geq 3$) and get R_i, update G_i to G_{i+1}
end for
Output G_n

References (note this is the full reference list - not all are cited in the supplementary information

1. Hajduk, P. J. & Greer, J. A decade of fragment-based drug design: strategic advances and lessons learned. *Nature reviews Drug discovery* **6**, 211–219 (2007).
2. Lawson, A. D. Antibody-enabled small-molecule drug discovery. *Nature Reviews Drug Discovery* **11**, 519–525 (2012).
3. Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* **4**, 279–287 (2022).
4. Freeze, J. G., Kelly, H. R. & Batista, V. S. Search for catalysts by inverse design: artificial intelligence, mountain climbers, and alchemists. *Chemical reviews* **119**, 6595–6612 (2019).
5. Gómez-Bombarelli, R. *et al.* Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature materials* **15**, 1120–1127 (2016).
6. Xu, R.-P., Li, Y.-Q. & Tang, J.-X. Recent advances in flexible organic light-emitting diodes. *Journal of Materials Chemistry C* **4**, 9116–9142 (2016).
7. Polishchuk, P. G., Madzhidov, T. I. & Varnek, A. Estimation of the size of drug-like chemical space based on gdb-17 data. *Journal of computer-aided molecular design* **27**, 675–679 (2013).
8. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
9. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery* **18**, 463–477 (2019).
10. Ekins, S. *et al.* Exploiting machine learning for end-to-end drug discovery and development. *Nature materials* **18**, 435–441 (2019).
11. von Lilienfeld, O. A., Müller, K.-R. & Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nature Reviews Chemistry* **4**, 347–358 (2020).
12. Westermayr, J., Gastegger, M., Schütt, K. T. & Maurer, R. J. Perspective on integrating machine learning into computational chemistry and materials science. *The Journal of Chemical Physics* **154**, 230903 (2021).
13. Zhang, Z., Liu, Q., Wang, H., Lu, C. & Lee, C.-K. Motif-based graph self-supervised learning for molecular property prediction. *Advances in Neural Information Processing Systems* **34**, 15870–15882 (2021).
14. Ceriotti, M., Clementi, C. & Anatole von Lilienfeld, O. Machine learning meets chemical physics (2021).
15. Keith, J. A. *et al.* Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chemical reviews* **121**, 9816–9872 (2021).
16. Fang, X. *et al.* Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence* **4**, 127–134 (2022).

17. Wang, D. *et al.* Efficient sampling of high-dimensional free energy landscapes using adaptive reinforced dynamics. *Nature Computational Science* **2**, 20–29 (2022).

18. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nature Biotechnology* 1–8 (2023).

19. Kusner, M. J., Paige, B. & Hernández-Lobato, J. M. Grammar variational autoencoder. In *International conference on machine learning*, 1945–1954 (PMLR, 2017).

20. Gómez-Bombarelli, R. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science* **4**, 268–276 (2018).

21. You, J., Liu, B., Ying, Z., Pande, V. & Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in neural information processing systems*, 6410–6421 (2018).

22. Griffiths, R.-R. & Hernández-Lobato, J. M. Constrained bayesian optimization for automatic chemical design using variational autoencoders. *Chemical science* **11**, 577–586 (2020).

23. Shi, C. *et al.* Graphaf: a flow-based autoregressive model for molecular graph generation. *International Conference on Learning Representations* (2020).

24. Gebauer, N., Gastegger, M. & Schütt, K. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. In *Advances in Neural Information Processing Systems*, 7566–7578 (2019).

25. Wang, J. *et al.* Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning. *Nature Machine Intelligence* **3**, 914–922 (2021).

26. Gebauer, N. W., Gastegger, M., Hessmann, S. S., Müller, K.-R. & Schütt, K. T. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature communications* **13**, 1–11 (2022).

27. Zhang, Z., Min, Y., Zheng, S. & Liu, Q. Molecule generation for target protein binding with structural motifs. In *The Eleventh International Conference on Learning Representations* (2022).

28. Zhang, Z. & Liu, Q. Learning subpocket prototypes for generalizable structure-based drug design. *ICML* (2023).

29. Ma, T., Chen, J. & Xiao, C. Constrained generation of semantically valid graphs via regularizing variational autoencoders. *Advances in Neural Information Processing Systems* **31** (2018).

30. De Cao, N. & Kipf, T. Molgan: An implicit generative model for small molecular graphs. *ICML 2018 workshop on Theoretical Foundations and Applications of Deep Generative Models* (2018).

31. Zang, C. & Wang, F. Moflow: an invertible flow model for generating molecular graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 617–626 (2020).

32. Madhawa, K., Ishiguro, K., Nakago, K. & Abe, M. Graphnvp: An invertible flow model for generating molecular graphs. *arXiv preprint arXiv:1905.11600* (2019).

33. Luo, Y., Yan, K. & Ji, S. Graphdf: A discrete flow model for molecular graph generation. In *International Conference on Machine Learning*, 7192–7203 (PMLR, 2021).

34. Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, 2323–2332 (PMLR, 2018).

35. Jin, W., Barzilay, R. & Jaakkola, T. Hierarchical generation of molecular graphs using structural motifs. In *ICML*, 4839–4848 (PMLR, 2020).

36. Ganea, O. *et al.* Geomol: Torsional geometric generation of molecular 3d conformer ensembles. *Advances in Neural Information Processing Systems* **34** (2021).

37. Xu, M. *et al.* An end-to-end framework for molecular conformation generation via bilevel programming. In *International Conference on Machine Learning*, 11537–11547 (PMLR, 2021).

38. Shi, C., Luo, S., Xu, M. & Tang, J. Learning gradient fields for molecular conformation generation. In *International Conference on Machine Learning*, 9558–9568 (PMLR, 2021).

39. Liu, S. *et al.* Pre-training molecular graph representation with 3d geometry. *International Conference on Learning Representations* (2022).

40. Mahmood, O., Mansimov, E., Bonneau, R. & Cho, K. Masked graph modeling for molecule generation. *Nature communications* **12**, 1–12 (2021).

41. Hoffmann, M. & Noé, F. Generating valid euclidean distance matrices. *arXiv preprint arXiv:1910.03131* (2019).

42. Hoogeboom, E., Satorras, V. G., Vignac, C. & Welling, M. Equivariant diffusion for molecule generation in 3d. *International Conference on Machine Learning* (2022).

43. Luo, Y. & Ji, S. An autoregressive flow model for 3d molecular geometry generation from scratch. In *International Conference on Learning Representations* (2021).

44. Luo, S., Guan, J., Ma, J. & Peng, J. A 3d generative model for structure-based drug design. *Advances in Neural Information Processing Systems* **34** (2021).

45. Méndez-Lucio, O., Ahmad, M., del Rio-Chanona, E. A. & Wegner, J. K. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nature Machine Intelligence* **3**, 1033–1039 (2021).

46. Liu, M., Luo, Y., Uchino, K., Maruhashi, K. & Ji, S. Generating 3d molecules for target protein binding. *International Conference on Machine Learning* (2022).

47. Peng, X. *et al.* Pocket2mol: Efficient molecular sampling based on 3d protein pockets. *International Conference on Machine Learning* (2022).

48. Vignac, C., Osman, N., Toni, L. & Frossard, P. Midi: Mixed graph and 3d denoising diffusion for molecule generation. *arXiv preprint arXiv:2302.09048* (2023).

49. Liu, Y. *et al.* Spherical message passing for 3d graph networks. *International Conference on Learning Representations* (2022).

50. Satorras, V. G., Hoogeboom, E., Fuchs, F. B., Posner, I. & Welling, M. E (n) equivariant normalizing flows. *NeurIPS* (2021).

51. Grosnit, A. *et al.* High-dimensional bayesian optimisation with variational autoencoders and deep metric learning. *arXiv preprint arXiv:2106.03609* (2021).

52. Notin, P., Hernández-Lobato, J. M. & Gal, Y. Improving black-box optimization in vae latent space using decoder uncertainty. *Advances in Neural Information Processing Systems* **34**, 802–814 (2021).

53. Maus, N. *et al.* Local latent space bayesian optimization over structured inputs. *Advances in Neural Information Processing Systems* **35**, 34505–34518 (2022).

54. Maus, N., Wu, K., Eriksson, D. & Gardner, J. Discovering many diverse solutions with bayesian optimization. *arXiv preprint arXiv:2210.10953* (2022).

55. Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S. & Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research* **22**, 1–64 (2021).

56. Dinh, L., Krueger, D. & Bengio, Y. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516* (2014).

57. Dinh, L., Sohl-Dickstein, J. & Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803* (2016).

58. O'Boyle, N. M. *et al.* Open babel: An open chemical toolbox. *Journal of cheminformatics* **3**, 1–14 (2011).

59. Kim, Y. & Kim, W. Y. Universal structure conversion method for organic molecules: from atomic connectivity to three-dimensional geometry. *Bulletin of the Korean Chemical Society* **36**, 1769–1777 (2015).

60. Garcia Satorras, V., Hoogeboom, E., Fuchs, F., Posner, I. & Welling, M. E (n) equivariant normalizing flows. *Advances in Neural Information Processing Systems* **34** (2021).

61. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data* **1**, 1–7 (2014).

62. Sun, Q. *et al.* Pyscf: the python-based simulations of chemistry framework. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **8**, e1340 (2018).

63. Sun, Q. *et al.* Recent developments in the pyscf program package. *The Journal of chemical physics* **153**, 024109 (2020).

64. Anderson, A. C. The process of structure-based drug design. *Chemistry & biology* **10**, 787–797 (2003).

65. Tripathi, A. & Bankaitis, V. A. Molecular docking: from lock and key to combination lock. *Journal of molecular medicine and clinical applications* **2** (2017).

66. Alon, A. *et al.* Structures of the σ2 receptor enable docking for bioactive ligand discovery. *Nature* **600**, 759–764 (2021).

67. Francoeur, P. G. *et al.* Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of chemical information and modeling* **60**, 4200–4215 (2020).

68. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature chemistry* **4**, 90–98 (2012).

69. Trott, O. & Olson, A. J. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **31**, 455–461 (2010).

70. Alhossary, A., Handoko, S. D., Mu, Y. & Kwoh, C.-K. Fast, accurate, and reliable molecular docking with quickvina 2. *Bioinformatics* **31**, 2214–2216 (2015).

71. Landrum, G. *et al.* Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **8** (2013).

72. Bento, A. P. *et al.* An open source chemical structure curation pipeline using rdkit. *Journal of Cheminformatics* **12**, 1–16 (2020).

73. Rappé, A. K., Casewit, C. J., Colwell, K., Goddard III, W. A. & Skiff, W. M. Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American chemical society* **114**, 10024–10035 (1992).

74. Ragoza, M., Masuda, T. & Koes, D. R. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chemical science* **13**, 2701–2713 (2022).

75. Axelrod, S. & Gomez-Bombarelli, R. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data* **9**, 1–14 (2022).

76. Peng, X., Guan, J., Liu, Q. & Ma, J. Moldiff: Addressing the atom-bond inconsistency problem in 3d molecule diffusion generation. *arXiv preprint arXiv:2305.07508* (2023).

77. Steinegger, M. & Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* **35**, 1026–1028 (2017).

78. Simm, G., Pinsler, R. & Hernández-Lobato, J. M. Reinforcement learning for molecular design guided by quantum mechanics. In *International Conference on Machine Learning*, 8959–8969 (PMLR, 2020).

79. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

80. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

81. Cowen-Rivers, A. I. *et al.* Hebo: pushing the limits of sample-efficient hyper-parameter optimisation. *Journal of Artificial Intelligence Research* **74**, 1269–1349 (2022).

82. Ghose, A. K., Viswanadhan, V. N. & Wendoloski, J. J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. a qualitative and quantitative characterization of known drug databases. *Journal of combinatorial chemistry* **1**, 55–68 (1999).

83. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews* **64**, 4–17 (2012).

84. Bajusz, D., Rácz, A. & Héberger, K. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics* **7**, 1–13 (2015).

85. Xu, M. *et al.* Geodiff: A geometric diffusion model for molecular conformation generation. *ICLR* (2022).

86. Halgren, T. A. Merck molecular force field. v. extension of mmff94 using experimental data, additional computational data, and empirical rules. *Journal of Computational Chemistry* **17**, 616–641 (1996).

87. Schütt, K. *et al.* Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in neural information processing systems*, 991–1001 (2017).

88. Schutt, K. *et al.* Schnetpack: A deep learning toolbox for atomistic systems. *Journal of chemical theory and computation* **15**, 448–455 (2018).