nature machine intelligence

MATTERS ARISING

https://doi.org/10.1038/s42256-020-00268-w

Reusability report: Designing organic photoelectronic molecules with descriptor conditional recurrent neural networks

Somesh Mohapatra, Tzuhsiung Yang and Rafael Gómez-Bombarelli ✉

ARISING FROM P.-C. Kotsias et al. *Nature Machine Intelligence* https://doi.org/10.1038/s42256-020-0174-5 (2020)

Deep generative models can be trained on unlabelled chemical data to design novel molecules, but it's challenging to harness the creativity of such models to finding optimal molecules[1]. In their recent work[2], Bjerrum and colleagues present a generative framework based on conditional recurrent neural networks (cRNNs) to translate from a desired property to a string-based chemical representation in the context of drug design. Here we replicate the approach on an unrelated chemical space by designing organic photoelectronic molecules (OPMs) with properties outside the training data. The primary application in the original work was a classification task (identifying active molecules) whereas the task here is to propose molecules with continuous properties close to target values.

The cRNN generative framework from ref. [2] can sample novel molecules conditioned on attributes such as structural fingerprints[3] or properties. Briefly, the model is trained to reproduce molecules by using their attributes to set the initial state of the RNN. At inference time, the desired molecular properties or molecular fingerprints are given as inputs to the cRNN and steer the stochastic generation of molecules. The approach thus aims to constrain the breadth of earlier RNN approaches through stronger supervision[4–6]. Estimates of the negative log-likelihood (NLL) of sampling given molecule allow interrogating the model in new ways. In ref. [2] molecular property models are trained on data labelled with inexpensive simulations. Transfer learning was then applied to adapt the model to the specific task of generating molecules that bind a particular target protein using a smaller set of labelled data. Because the approach operates on text-based molecular representations (specifically using the simplified molecular-input line-entry system, SMILES) it requires data augmentation to avoid pitfalls arising from atom indexing and non-canonical SMILES.

Generating OPMs using cRNNs

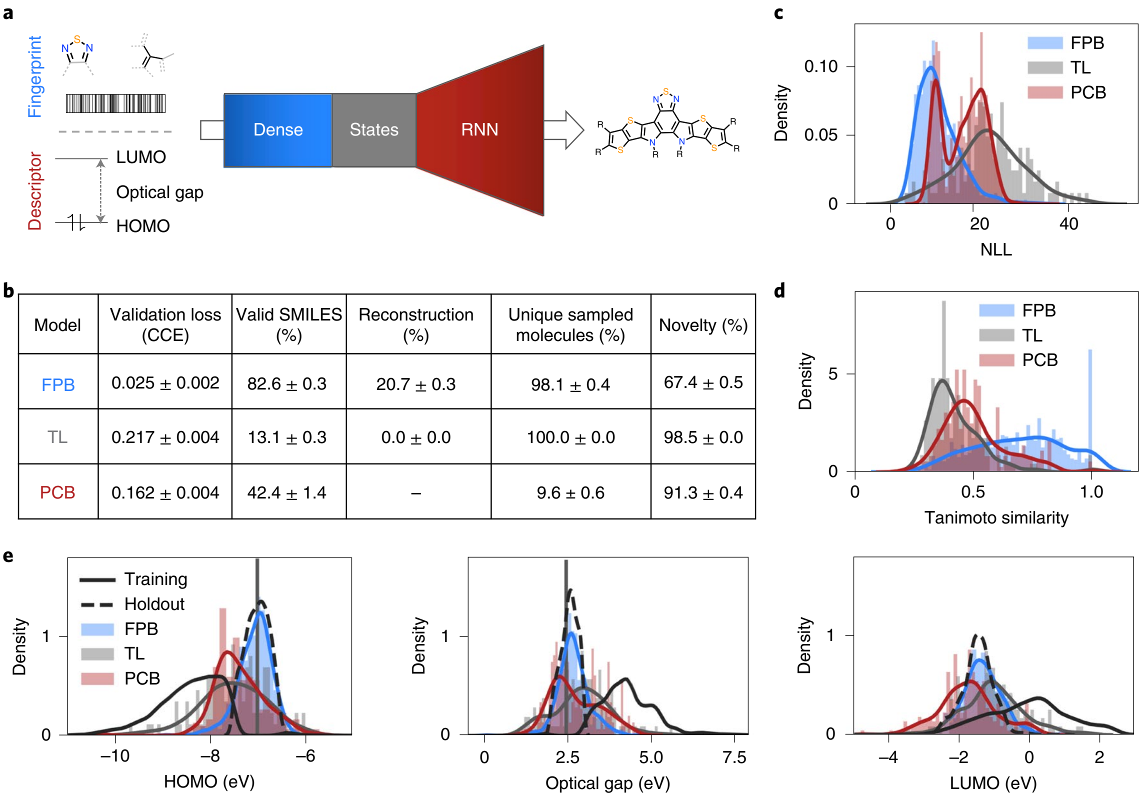
OPMs have a variety of applications and designing optimal OPMs is highly desirable for technologies such as transistors, displays[7] and solar cells[8]. While not as chemically diverse as small-molecule drugs, the space of potential OPMs is vast. OPMs typically contain conjugated heterocycles and have sizes in the tens of heavy atoms, and thus they span a very distinct design space (Supplementary Section 5). Key attributes for OPMs are their electronic and optical properties, which can be quantified as the energies of their electron-filled highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO), which are related to their ability to transport holes and electrons, respectively, and the energy needed to promote one electron from the occupied to the unoccupied orbitals by absorbing light (optical gap). These energy levels can be simulated with reasonable accuracy through density functional theory (DFT), which allows obtaining property labels to train the cRNN generative models. Typical calculated values for OPMs in electron volts (eV) are $-10 < \mathrm{HOMO} < -6$; $-4 < \mathrm{LUMO} < 2$; and $1 < \text{optical gap} < 5$.

We tested the ability of cRNN models to generate OPMs with desired properties. Unlike the work in ref. [2] the desired properties are continuous and not categorical. As training data, we utilized chemical structures of molecules extracted from the literature, US patents and combinatorially generated derivatives (Supplementary Sections 1,2). In total, the available data summed to about 172,000 molecules of which 14,800 had been labelled with HOMO, LUMO and optical gap using DFT calculations (Supplementary Sections 3,4).

Following the original work and codebase[9], with minimal modifications to handle larger molecules as well as different descriptors and optimization of learning rate, we trained and validated three different models using the mutually exclusive unlabelled (157,665), labelled (13,616), and seed (1,129) datasets described in Supplementary Section 4. A fingerprint-based model was trained on unlabelled data (FPB). A sequentially transfer-learnt (TL) model was trained starting from the FPB weights using the labelled dataset. A descriptor-based (PCB) model was trained with HOMO, LUMO and optical gap labels as descriptor inputs (Fig. 1a, Supplementary Section 6 for details on other descriptor models). The FPB model was trained on the unlabelled dataset, excluding all molecules that were present in the labelled dataset; TL and PCB models were trained on the labelled dataset, with molecules having the desired properties being held out (seed dataset of 1,129 molecules with $-7.5 < \mathrm{HOMO} < -6.5$ or $2 < \text{optical gap} < 3$). The validation loss for FPB (Fig. 1b) is the best of the three, suggesting that performance at generative tasks is impacted by the lower data size of 13,600.

We evaluated the ability of the fingerprint models to (re)construct held-out molecules with desired attributes, namely a HOMO of −7.0 eV, which can be related to efficient charge conductivity, and a 2.5 eV optical gap, at which the solar flux has the highest intensity[10] (Supplementary Section 7). The FPB model performed better at reconstruction, and slightly worse at novelty (Fig. 1b). The mean NLL scores for reproducing the desired molecules and mean Tanimoto similarity scores with the training data (Fig. 1c,d) suggest that the FPB model has captured the space of OPMs fairly well, and can produce a variety of novel, valid, OPMs with desired properties if seeded with high-performance candidates. The TL model suffers from some degree of forgetting. Since the labelled and unlabelled

Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. ✉e-mail: rafagb@mit.edu



Model	Validation loss (CCE)	Valid SMILES (%)	Reconstruction (%)	Unique sampled molecules (%)	Novelty (%)
FPB	0.025 ± 0.002	82.6 ± 0.3	20.7 ± 0.3	98.1 ± 0.4	67.4 ± 0.5
TL	0.217 ± 0.004	13.1 ± 0.3	0.0 ± 0.0	100.0 ± 0.0	98.5 ± 0.0
PCB	0.162 ± 0.004	42.4 ± 1.4	–	9.6 ± 0.6	91.3 ± 0.4

Fig. 1 | Repurposing cRNN generates novel OPMs. a, The model framework used in the generation of OPMs from cRNN. **b**, Training and sampling metrics for different models: FPB (trained on unlabelled dataset of 157,665 molecules), TL (FPB transfer learnt using $N=13{,}616$ labelled molecules) and PCB (descriptor-based model, $N = 13{,}616$). Validation loss (categorical cross-entropy, CCE) is presented for the best performing model in 1,000 epochs. All sampling metrics were acquired during generation of molecules with desired properties (Supplementary Section 7). Valid SMILES is the percentage of strings that can be processed by RDKit[21]. Reconstruction is the percentage of times when the sampled molecule is the same as the seed (N/A for the descriptor-based model). Unique sampled molecules is the percentage of novel molecules sampled. Novelty is the percentage of sampled molecules that are not in the respective training datasets. **c**, Average NLL of producing molecules obtained in the design experiments. The FPB model shows much sharper distribution. PCB shows two peaks, one of which can be attributed to mode collapse. TL has the poorest average, and is smoother. **d**, The distribution of maximum Tanimoto similarity scores of the sampled molecules against molecules in the training dataset. FPB either reconstructs or weakly mutates, but TL and PCB have broader distributions. **e**, The distribution of molecular properties of training, holdout and sampled molecules. The vertical lines represent the desired HOMO and optical-gap values.

data belong to the same chemical space, this performance suggests that transfer to new chemical spaces would require substantially more training data and strategies such as semisupervised training with both the labelled and unlabelled sets.

We then tested the ability of the PCB model to produce desired molecules without a need for seed molecules. As inputs, we used 10,000 property combinations chosen at random within the range of designed properties (Supplementary Section 7). The sampling resulted in 42% valid SMILES and 91% novel molecules (Fig. 1b). However, only 10% of valid SMILES were unique, suggestive of substantial mode collapse of the PCB model compared with the more data-rich FPB and equally data-poor TL models.

DFT calculations were then performed to validate the properties of the sampled molecules (Fig. 1e). All models generated molecules with properties much closer to the desired values than to the labelled training data. Respectively, the FPB, TL and PCB models produced mean HOMO of −7.07, −7.50 and −7.41 eV (s.d.: 0.41, 0.70 and 0.67) and mean optical gap of 2.72, 2.97 and 2.78 eV (s.d.: 0.49, 0.85 and 0.73).

Benchmarking cRNN models against a classical baseline

We benchmarked the cRNN models with a simpler graph-based genetic algorithm (GB-GA) approach that has recently shown very good performance in molecular optimization[11]. The GB-GA model slightly underperformed compared with the cRNN models for finding molecules with desired properties (HOMO, optical gap; Fig. 2; Supplementary Section 8). Both approaches produced a substantial fraction of unrealistic molecules. In the case of the cRNN models, these were typically due to errors in the character-wise decoding of SMILES (missing a cycle closure, unfeasible four- and eight-membered aromatic rings). Such errors can, in principle, be addressed by better embedding the chemical space with more training data and more powerful models. The GB-GA approach relies on hand-selected rules, which are applied without awareness of the chemical context and tend to result in incompatible chemical groups and unreasonable functionalization. This could only be addressed with additional hard-coded rules about chemical feasibility. Those rules are lacking, however, which is one of the key drivers for the development of generative models for molecules.

Discussion

The cRNN approach of ref. [2] was found to be generally applicable for the design of OPMs and agree well with the original work, despite the fact that most OPMs are much larger than typical small-molecule drugs, which makes the generation of valid SMILES and reconstruction harder tasks. In addition, organic-photoelectronic properties are global and non-additive, unlike the cheminformatics descriptors used to train supervised generative models for drug-like

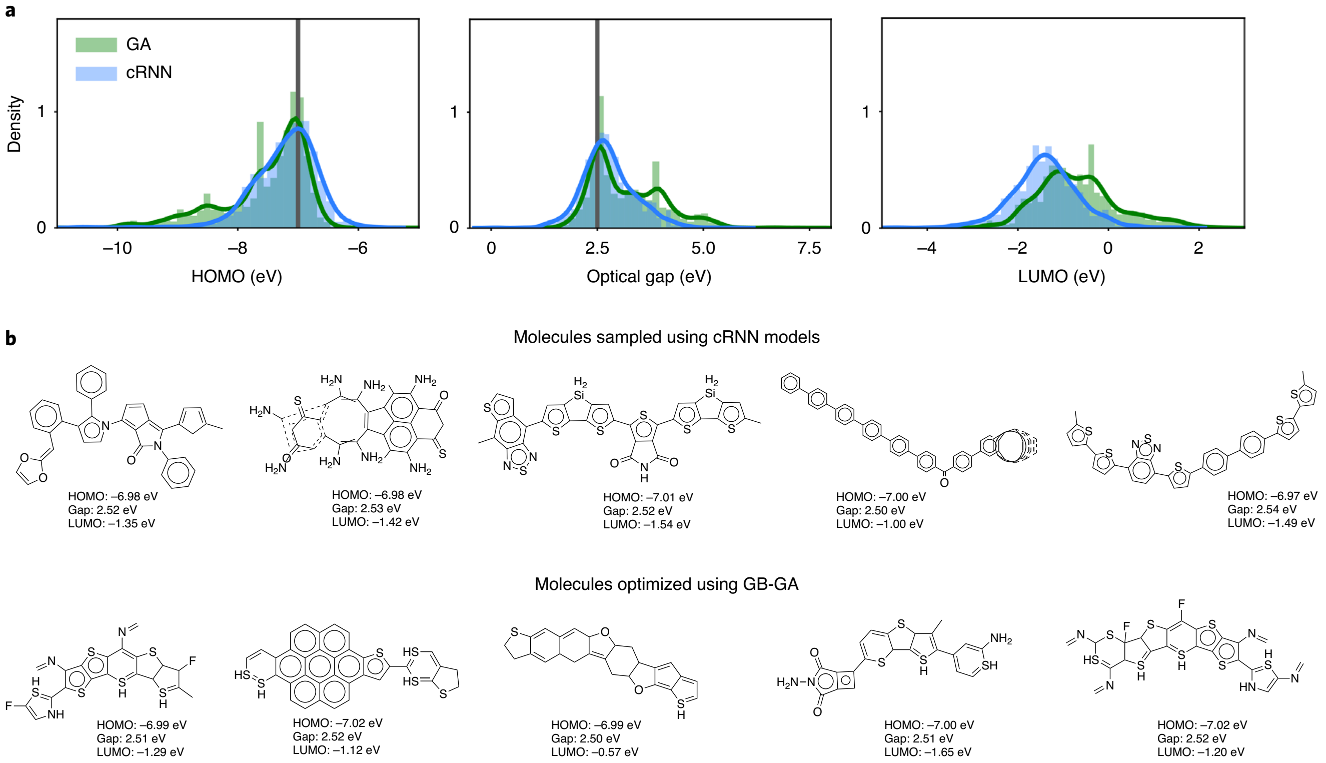


Fig. 2 | Benchmarking cRNN models against a GB-GA baseline. a, Comparison of distributions of properties for molecules sampled using cRNN models, and optimized using a GB-GA approach. **b**, Comparison of five random molecules sampled using cRNN models, and optimized using a GB-GA approach. From the ten molecules shown, only the first and fifth (from left) cRNN molecules look realistic.

molecules (Supplementary Section 5). Finally, the size of labelled datasets is typically lower in OPMs since the DFT calculations that produce HOMO, LUMO and optical gap values are much more expensive than the cheminformatic methods that generate descriptors for drug-like molecules (log *P*, Lipinski's rule and so on) used in most works on generative models for molecular optimization. High-throughput and combinatorial experiments to produce reference datasets are also more common in small-molecule drugs than organic photoelectronics.

Where this work disagrees with ref. [2], we found data availability to be a key driving factor. The FPB model outperformed the other two models at generation of molecules with desired properties, by essentially adding chemical noise to seed molecules with good properties and leveraging its reconstruction accuracy. However, the FPB model is limited in its ability to invent molecules with never-seen performance, as the sampling approach requires the fingerprints of seed molecules that already have near-desired performance. The FPB clearly outperformed TL, which was not the case in ref. [2]. This is due to the smaller-sized labelled dataset, which results in catastrophic forgetting. Although semisupervised approaches could address this, they are not applicable in this case, since one dataset is an excise of the other. Because both belong to the same manifold of chemical space, semisupervised techniques would be unrealistically favourable. This sharp loss in performance suggests that some further tuning of the sequential TL strategy is needed if transferring to a labelled training dataset of size ~10^4 or less. Needing no seed molecules, the PCB strategy was capable of inferring molecules with desired properties despite their absence from the training data, and thus learned how to navigate structure–property relationships. Although it suffered from more intense mode collapse and lower validity and reconstruction than the FPB model, it clearly outperformed the TL model. This can be rationalized from the much lower dimensionality of its input space, 3 descriptor variables compared with 2,048 bits.

In summary, all the models produced diverse OPMs with desired properties and effectively moved away from the property distribution of the training data. The key limiting factor was data availability as evidenced by the both the TL and PCB models.

Future directions

We identify two avenues of interest for further development of the cRNN approach for molecular design. One is more efficient TL strategies, such as freezing certain weights, particularly given the scarcity of labelled molecular data to train descriptor models. Innovations such as regret minimization show promise for molecular design[12]. The other avenue is avoiding the generation of text-based molecular representations, which are not permutationally invariant and rely on a complex grammar[13]. The use of other string representations[14], grammar-[15] or syntax-based[16] approaches may improve validity and novelty in low-data regimes. Similarly, for large molecules stack-augmented memory units[17], or nested architectures[18] or transformer models[19] may be better at capturing the long-range relationships in the SMILES grammar.

Data availability

The chemical structures and labels used for training and validation of the supervised and unsupervised models, with the exception of 684 proprietary molecules, are available at https://github.com/learningmatter-mit/Deep-Drug-Coder[20].

Code availability

The code used in this paper is available at https://github.com/learningmatter-mit/Deep-Drug-Coder.

Received: 21 June 2020; Accepted: 27 October 2020;
Published online: 10 December 2020

References

1. Schwalbe-Koda, D. & Gómez-Bombarelli, R. in *Lecture Notes in Physics* Vol. 968 (eds Schütt, K. T. et al.) 445–467 (Springer, 2020).
2. Kotsias, P.-C. et al. Direct steering of de novo molecular generation using descriptor conditional recurrent neural networks (cRNNs). *Nat. Mach. Intell.* **2**, 254–265 (2020).
3. Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chem. Doc.* **5**, 107–113 (1965).
4. Segler, M. H. S., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.* **4**, 120–131 (2018).
5. Arús-Pous, J. et al. Exploring the GDB-13 chemical space using deep generative models. *J. Cheminform.* **11**, 20 (2019).
6. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**, eaap7885 (2018).
7. Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
8. Hachmann, J. et al. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry – the Harvard Clean Energy Project. *Energy Environ. Sci.* **7**, 698–704 (2014).
9. Kotsias, P. & Bjerrum, E. J. Deep-Drug-Coder v1.0.0 https://doi.org/10.5281/zenodo.3739063 (accessed 15 May 2020).
10. Gueymard, C. A. The sun's total and spectral irradiance for solar energy applications and solar radiation models. *Sol. Energy* **76**, 423–453 (2004).
11. Jensen, J. H. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem. Sci.* **10**, 3567–3572 (2019).
12. Jin, W., Barzilay, R. & Jaakkola, T. Domain extrapolation via regret minimization. Preprint at https://arxiv.org/abs/2006.03908 (2020).
13. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
14. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* **1**, 045024 (2020).
15. Kusner, M. J., Paige, B. & Hernández-Lobato, J. M. Grammar variational autoencoder. In *Proc. 34th International Conference on Machine Learning* (eds Precup, D. & Teh, Y. W.) 1945–1954 (2017).
16. Dai, H., Tian, Y., Dai, B., Skiena, S. & Song, L. Syntax-directed variational autoencoder for molecule generation. In *Proc. International Conference on Learning Representations* (ICLR, 2018).
17. Joulin, A. & Mikolov, T. Inferring algorithmic patterns with stack-augmented recurrent nets. In *Advances in Neural Information Processing Systems* (2015).
18. Moniz, J. R. A. & Krueger, D. Nested LSTMs. In *Proc. Asian Conference on Machine Learning* (PMLR, 2017).
19. Maziarka, Ł. et al. Molecule attention transformer. Preprint at https://arxiv.org/abs/2002.08264 (2020).
20. Mohapatra, S., Yang, T. & Gomez-Bombarelli, R. OPM-cRNN v0.1-OPM https://doi.org/10.5281/zenodo.4073289 (2020).
21. Landrum, G. RDKit: Open-source cheminformatics v2018.09.1 https://www.rdkit.org/docs/index.html (2006).

Acknowledgements

We acknowledge Sumitomo Chemical for providing financial support for this work.

Author contributions

R.G.-B. supervised the research, and planned the project with contributions from S.M. S.M. trained and analysed the machine learning models. T.Y. ran the DFT calculations with contributions from R.G.-B. All authors contributed to the writing of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-020-00268-w.

Correspondence and requests for materials should be addressed to R.G.-B.

Peer review information *Nature Machine Intelligence* thanks Olexandr Isayev, Connor Coley and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.