Advances in Condensed Matter Physics 凝聚态物理学进展, 2020, 9(2), 11-19
Published Online May 2020 in Hans. http://www.hanspub.org/journal/cmp
https://doi.org/10.12677/cmp.2020.92002

Hans汉斯

Application of Machine Learning in Material Property Prediction

Lei Ti[1], Sixuan Wu[2], Bin Li[2,3*], Yuheng Shi[2], Ziwan Song[1], Zilong Miao[1], Zhixiang Shi[4]

[1]College of Electronic and Optical Engineering, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu
[2]School of Science, Nanjing University of Posts and Telecommunications, Nanjing Jiangsu
[3]New Energy Technology of Jiangsu Province, Nanjing Jiangsu
[4]Department of Physics, Southeast University, Nanjing Jiangsu
Email: *libin@njupt.edu.cn

Received: Mar. 24th, 2020; accepted: Apr. 7th, 2020; published: Apr. 14th, 2020

Abstract

The prediction of enthalpy of formation of the crystal is studied by the machine learning algorithm. Based on the relationship between the learning data of the deep neural network (DNN) model composed of multiple neural layers, the prediction of enthalpy of formation of the deep learning model in the material junction is studied and discussed in depth. By learning the enthalpy of formation parameters of 275,778 compounds in the open quantum materials database (OQMD), a deep learning multi-layer fully connected network was established to predict the enthalpy of formation of unknown crystal materials. The accuracy of the optimized prediction model reached 0.075 eV/atom, which reached the computational accuracy of the quantum mechanics software.

Keywords

Crystal Structure, Machine Learning, High-Throughput Calculation

机器学习在材料性质预测中的运用

提　磊[1]，吴思璇[2]，李　斌[2,3*]，石宇衡[2]，宋紫菀[1]，缪子隆[1]，施智祥[4]

[1]南京邮电大学电子与光学工程学院，江苏　南京
[2]南京邮电大学理学院，江苏　南京
[3]江苏省新能源技术工程实验室，江苏　南京
[4]东南大学物理学院，江苏　南京
Email: *libin@njupt.edu.cn

*通讯作者。

文章引用: 提磊, 吴思璇, 李斌, 石宇衡, 宋紫菀, 缪子隆, 施智祥. 机器学习在材料性质预测中的运用[J]. 凝聚态物理学进展, 2020, 9(2): 11-19. DOI: 10.12677/cmp.2020.92002

收稿日期：2020年3月24日；录用日期：2020年4月7日；发布日期：2020年4月14日

摘　要

本文采用机器学习算法对晶体的生成焓的预测进行了研究。利用由多个神经层组成的深度神经网络(DNN)模型学习数据间的关系，对深度学习模型在材料结生成焓的预测做了深入的研究和讨论。通过对开放量子材料数据库(OQMD)中275,778种化合物的生成焓参数进行学习，建立了深度学习多层全连接网络，并用来预测未知晶体材料的生成焓。优化后预测模型的精度达到了0.075 eV/atom，达到了量子力学软件的计算精度。

关键词

晶格结构，机器学习，高通量计算

1. 引言

近年来，第一性原理高通量计算(ab-initio high-throughput computational methods)在预测新材料和优化材料的属性等方面被证明是一种强大而且成功的方法。例如：多组分晶体、合金相图的成功预测[1] [2]，太阳能电池材料的电极透明度和电导率的优化，得到更好的电流－电压特性[3]。然而，大量的计算资源需求成为了高通量计算的瓶颈。当化合物可能的结构搜索空间变得很大，元胞的结构变得很复杂时，即使采用高效的 Kohn-Sham 密度泛函理论(KS-DFT)，所需的计算量对于有限的计算资源来说也是一项挑战。人们希望有一种更快的方式来预测新材料的物理属性，而无需求解 KS-DFT 方程。最近，机器学习(ML)技术逐渐被人们重视。对于预测新材料的物理属性研究方面，机器学习技术和传统的基于第一性原理计算的本质区别在于，机器学习摒弃了传统计算方法中通过求解多电子体系薛定谔方程来预测材料的物理属性，而是直接从大量已知的物理数据出发，通过多层次神经网络的运算，找到不同材料的不同物理属性的内在网络关系，从而迅速地得到未知新材料的物理属性。基于机器学习的计算效率很高，模型训练结束后，预测给定材料的特定属性通常仅需要几秒。此外，机器学习在预测分子特性，过渡态，表面反应等领域也有了较大发展，出现了一系列精准的模型[4] [5] [6] [7]。

深度神经网络(DNN)模型可以直接从输入表示中学习，例如文本的数字编码，图像的颜色像素等，无需研究人员考虑如何进一步描述数据[8]，从而省略了建立传统机器学习模型中所需的特征工程等人工步骤。由于这种优势，深度学习近些年在计算机科学领域获得了极大的关注，在计算机视觉[9]、语音识别[10]和文本处理领域[11]-[18]取得了突破性进展。尽管深度学习模型在上述应用中取得了巨大成功，但材料科学中深度学习的发展还处于早期阶段。本文采用深度学习快速预测晶体性质，利用由多个神经层组成的 DNN 模型学习数据间的关系，对深度学习模型在材料结生成焓的预测做了深入的研究和讨论。

2. 神经网络预测晶体生成焓

我们利用深度学习，直接从元素的组成中学习材料的属性，消除了当前需要手动特征工程的局限性。

它仅将元素组成作为输入，并利用网络自动捕获基本化学关系来预测生成焓。

人工神经网络由许多人工神经元连接而成。它们接受输入，然后根据输入调整自身的激活状态，并根据输入和激活状态产生输出。通常，神经元 j 除了从它的上级神经元接收到的输入 $p_j(t)$ 以外，它的状态还包含：自身的激活状态 $a_j(t)$、阈值 θ_j、激活函数 f 和输出函数 f_{out}。

激活函数 f 用于计算时刻 t 到 $t+1$ 时的激活状态，取决于 θ_j 和净输入 $p_j(t)$ 的关系：

$$a_j(t+1)=f\left(a_j(t),p_j(t),\theta_j\right) \tag{1}$$

神经元之间相互连接，每个连接将神经元 i 的输出传递给神经元 j 的输入。因此，i 是 j 的上级，j 是 i 的下级，连接会被分配一个权重 w_{ij}。因此，上级神经元的输出 $o_j(t)$ 计算输入到神经元 j 的输入 $p_j(t)$ 取决于其所有上级(有连接的)神经元的输出和对应连接的 w_{ij}：

$$p_j(t)=\sum_i o_j(t)w_{ij} \tag{2}$$

当需要添加一个偏差项时，方程变为：

$$p_j(t)=\sum_i o_j(t)w_{ij}+w_{0j} \tag{3}$$

其中 w_{0j} 为偏差。

训练一个神经网络本质上是在所有的可能的模型参数中(超空间中)选择能正确映射输入与期望的模型。有许多算法可以训练神经网络，其中绝大多数可以被视为是一种基于统计估计的优化理论。

绝大多数是采用反向传播来计算梯度下降。只需计算损失函数的梯度然后根据该梯度调节网络的参数，使结果更靠近最优处。反向传播的主要特征是迭代，循环更新网络参数，直到网络能够有效完成正在训练的任务，将输入准确的映射到期望的输出。反向传播的权重更新可以通过随机梯度下降使用以下等式完成：

$$w_{ij}(t+1)=w_{ij}(t)-\eta\frac{\partial C}{\partial w_{ij}}+\xi(t) \tag{4}$$

其中 η 是学习率，C 是损失函数，$\xi(t)$ 是一个随机数。损失函数的选择取决于学习类型和激活函数。

机器学习中的学习率 η 是一个超参数，它决定神经元新获取的信息在多大程度上覆盖旧信息。η 的选择很重要，因为较高的值会导致过强的变化，导致错过最小值，而过低的速度会不必要地减慢训练速度。η 应始终小于 1，否则网络将不会收敛，通常 $\eta\in[0.0001.0.4]$。η的最佳值因模型和数据而异，所以为了加速误差最小化，提高可靠性，η 通常在训练期间根据学习速率表或通过使用自适应学习速率而变化。目前，我们有很多优化函数可使用，比如 Adagrad，Adadelta，RMSprop，Adam 等，它们通常内置于深度学习库中。其中 RMSprop 算法基于 Rprop，是一个专为神经网络设计的优化算法，近年来在自适应学习方法中受到广泛关注。

最优问题的梯度的幅度可能变化很大，这时我们找到模型合适的学习率就变得困难。如果使用全批量学习模式，只能计算模型的平均梯度来确定学习率。这种方式对于高梯度的鞍点或极值点位置没什么问题，因为只要我们有足够的迭代次数，即使每一步的步进距离很小，我们最终还是能到达最优点。然而，如果模型的最优点附近梯度很小，很可能被越过。Rprop 试图让学习率根据模型自动变化，在梯度小的区域变小，在梯度大的区域变大。该方法结合了梯度和权重模型，每次迭代都会根据特定的权重自动调整步长。首先，查看最后两步的梯度，如果它们的符号相同，证明搜索的方向正确，可以轻微的增加步长；如果符号相反，说明搜索的步长过大了，跳过了一个局部最小，因此要减小步长。最后为步长设置上下限，这个上下限需要根据模型的细节和数据集来设定。

Rprop 的不足在于，当数据集非常大的时候，它不能执行小批量权重更新。例如：我们有 10 组数据，其中 9 个的梯度为 0.1，1 个为 1。我们希望这个梯度能相互抵消，但是 Rprop 算法将连续增加 9 次权重，导致权重更大，步长变得不可控，因此本文采用 RMSprop 算法。

RMSprop 的核心思想是保证权重是梯度的均方根。

$$\mathrm{E}\left[g^2\right]_t = \beta \mathrm{E}\left[g^2\right]_{t-1} + (1-\beta)\left(\frac{\delta C}{\delta w}\right)^2 \tag{5}$$

$$w_t = w_{t-1} - \frac{\eta}{\sqrt{\mathrm{E}\left[g^2\right]_t}} \frac{\delta C}{\delta w} \tag{6}$$

其中 $\mathrm{E}\left[g^2\right]_t$ 是平方梯度的平均，$\frac{\delta C}{\delta w}$ 是损失函数权重的导数，h 是学习率，β 是个参数。

为了找到最佳的模型，我们先从两层网络开始，进行训练，测试精确度，之后逐步增加网络层数和神经元数，直到精确度不再显著增加。在两个拥有不同神经元的层之间需要加入 dropout 层[19]，防止过拟合的发生[20]。在层数超过 6 层时，测试误差不再减小反而上升，因此我们认为此时的模型已经能学习到数据中必要的特征。进一步增加网络层数已经不能再提高精确度，但会显著增加模型训练的时间和过拟合的可能，所以我们的不再继续增加层数。我们还尝试了不同类型的激活函数，其中 ReLU 的表现最好[20]。

本文模型所使用的 275,778 种化合物数据全部来自 OQMD 数据库，OQMD 是一个广泛使用的高通量密度泛函(DFT)数据库，其中的数据包含了 DFT 计算过的晶体学参数，无机晶体结构数据库(ICSD)中生成焓的实验值等数据。我们选取每种成分的最低生成焓训练我们的预测模型，因为它们代表了最稳定的化合物，这使得我们的模型能够预测给定成分的基态结构的能量。275,778 种化合物数据中，6.3%的材料由双元素构成，81.4%的材料由三元素构成，12.26%的材料由四元素构成。我们需要将化学式转化为模型能识别的表示方式：采用一个固定长度的向量，里面记录着每种元素原子数的比例值(当元素没有出现在化合物中时为 0)，我们的多层神经网络拓扑结构如图 1 所示。

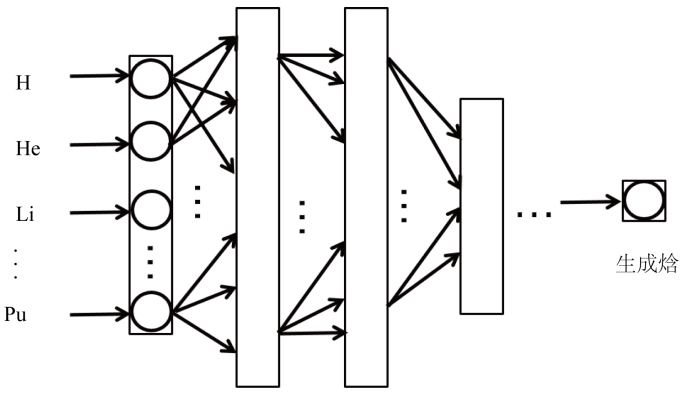

Figure 1. Multilayer neural network topology. Elements are arranged according to the number of protons, and the normalized proportional value of elements is taken as the input vector of the input layer.
图 1. 多层神经网络拓扑结构。将元素按质子数排列，将元素的归一化比例值作为输入层的输入向量

图 2 展示了实验过程中我们测试过的四种模型。标签 32 × 2 + 16 × 2 + 8 × 1 + 1 × 1 代表模型的架构为：两层 32 神经元网络加上两层 16 神经元网络加一层 8 神经元网络加一层 1 神经元网络(输出层)。由图

可知当神经元层数增加，神经元数量增加，模型的误差下降。128 × 2 + 64 × 2 + 32 × 1 + 1 × 1 结构的模型在我们所有的测试中误差是最低的。从图 3 我们可以看到，当我们继续增加神经元层，发现 256 × 2 + 128 × 2 + 64 × 2 + 32 × 1 + 1 × 1 结构出现了过拟合：即随着迭代次数的增加，训练集的误差(loss)在下降，但测试集的误差(val loss)不断上升。

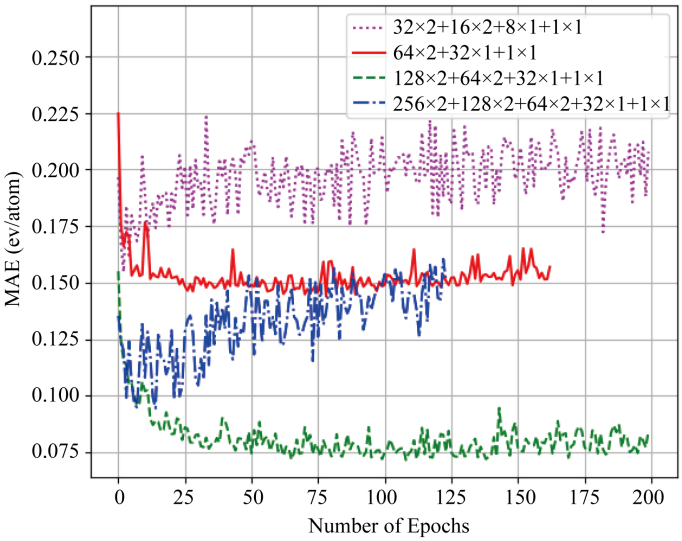

Figure 2. Relations between number of layers and error, the x-axis is the layer number model, and the y-axis is the error
图 2. 层数与误差的关系，横坐标为层数模型，纵坐标为误差

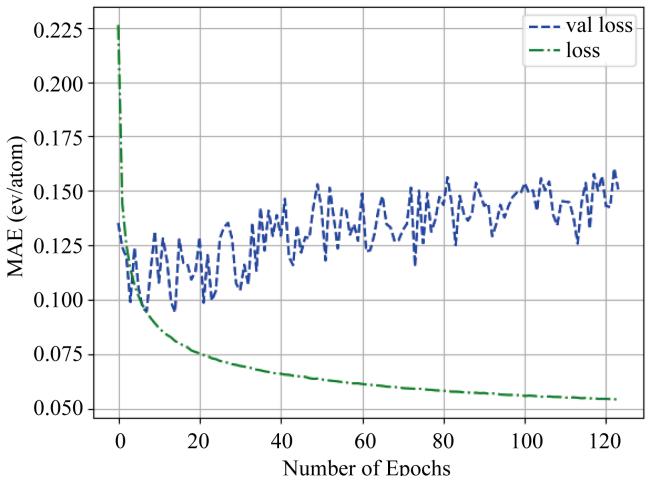

Figure 3. Rraining situation of 256 × 2 + 128 × 2 + 64 × 2 + 32 × 1 + 1 × 1 structure.
图 3. 256 × 2 + 128 × 2 + 64 × 2 + 32 × 1 + 1 × 1 结构的训练情况

我们所使用的学习模型的最终结构见表 1：

输入层为第 0 层，不同类型的全连接层和 dropout 的位置如表所示，dropout 层用于防止过度拟合，它们不计为单独的层，模型中我们使用 ReLU 作为激活函数。

深度学习模型在许多应用中取得了巨大成功，但通常这些是训练数据相对丰富的应用，但数据量并

非越多越好。一方面，复杂模型需要更多的训练数据，否则容易出现过拟合。另一方面，如果采用简单模型，即使给与大量的训练数据，也会因为模型无法描述复杂的关系而无法做出精准的映射，而且过量的数据会减慢训练速度。为了理解本深度学习模型的精度与数据集大小之间的关系，我们使用十折交叉验证的方法，比较了训练数据集大小对深度学习模型的准确性的影响。十折交叉验证是把样本数据随机分成 10 份，每次随机选择 9 份作为训练集，剩下的 1 份做测试集。训练结束后，重新随机选择 9 份再次进行训练，直至得到损失函数评估最优的模型和参数。本文中我们使用了四种大小的数据集分别进行研究：5000、10000、50000、275778。

Table 1. Structure of deep learning model
表 1. 深度学习模型结构

层类型	单元数	激活函数	位置(层)
全连接层	128	ReLU	1~2
dropout	128	--	--
全连接层	64	ReLU	3~4
全连接层	32	ReLU	5
全连接层	1	Linear	6

由图 4 结果可知，该深度学习模型在数据集越大时，模型误差越小，精度越高。这说明我们的模型足够复杂，在大量的数据中能捕捉到其内在关系。我们将所有数据的 90%作为训练集并经过多轮迭代后，模型逐渐收敛于某个最优解，见图 5。

图 6 是使用 25662 个元素组成的测试集对模型进行 10 折交叉验证的结果。在其左图中，散点分布于 $y = x$ 附近说明了深度学习模型的预测值和 DFT 计算值相接近。该模型预测的平均绝对值误差为 0.075 eV/atom，表现出了很高的准确性。如右图所示，90%的预测误差小于 0.180 eV/atom，80%的预测误差小于 0.080 eV/atom。

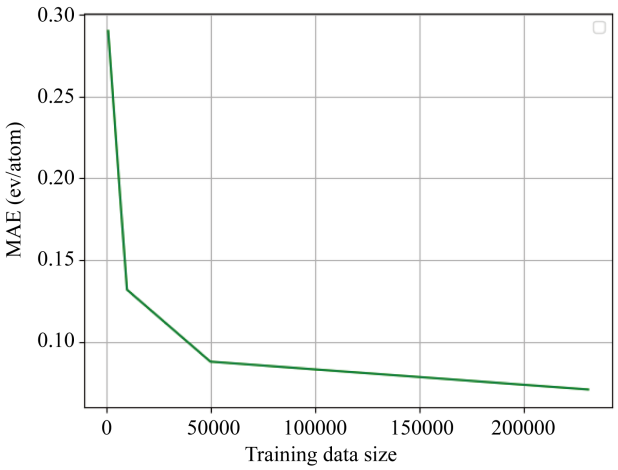

Figure 4. Relationship between dataset size and accuracy of deep learning model
图 4. 数据集大小与深度学习模型的准确性的关系

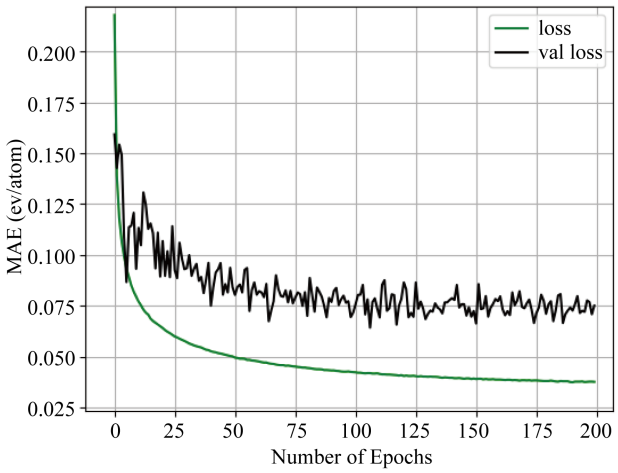

Figure 5. Performance of the model after several iterations. “loss” is the average absolute difference of the training set. “var loss” is the average absolute difference of the 10-fold cross-validation of the model with the test set consisting of 25,662 elements. The test set accuracy of the model converges to 0.075 eV/atom

图 5. 模型在经过多次跌迭代后的表现。其中 loss 是训练集的平均绝对差值。var loss 是使用 25662 个元素组成的测试集对模型进行 10 折交叉验证的平均绝对差值。模型的测试集精度收敛于 0.075 eV/atom

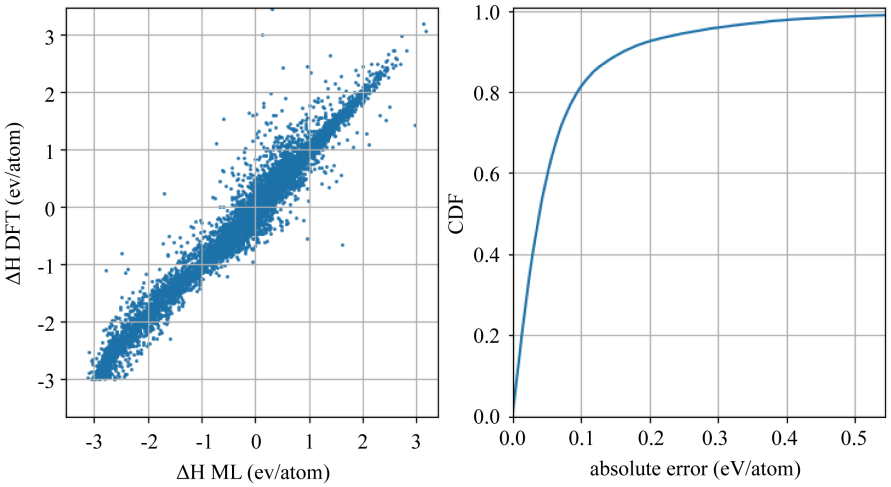

Figure 6. The results of a 10-fold cross-validation of the model using a test set of 25,662 elements. The *x*-axis of the left figure is the predictions of the neural network model, and the *y*-axis is the DFT calculation results. The right figure shows the cumulative distribution function (CDF)

图 6. 使用 25662 个元素组成的测试集对模型进行 10 折交叉验证的结果。左图的 *x* 轴是神经网络模型的预测，*y* 轴是 DFT 计算结果。右图为累积分布函数(CDF)

3. 结论

本文使用了机器学习算法在 OQMD 数据库的基础上建立了深度学习多层全连接网络，用来预测材料的生成焓。在优化后预测模型中，80%的预测误差小于 0.080 eV/atom，平均误差为 0.075 eV/atom，达到了 DFT 计算的精度，说明我们的模型可以用来预测未知材料化合物生成焓，之后我们将尝试将机器学习运用在其他材料属性的预测中。

致　谢

感谢南京大学固体微结构物理国家重点实验室 e-Science 中心的计算支持。

基金项目

国家自然科学基金面上项目(批准号：11674054)、国家自然科学基金青年科学基金(批准号：11504182)、南京邮电大学校级科研基金资助项目(批准号：NY219087，NY220038)和南京大学固体微结构物理国家重点实验室开放项目资助的课题(批准号：M32025)。

参考文献

[1] Curtarolo, S., Kolmogorov, A.N. and Cocks, F.H. (2005) High-Throughput *ab Initio* Analysis of the Bi-In, Bi-Mg, Bi-Sb, In-Mg, In-Sb, and Mg-Sb systems. *Calphad*, **29**, 155-161. https://doi.org/10.1016/j.calphad.2005.04.003

[2] Oganov, A.R. and Glass, C.W. (2006) Crystal Structure Prediction Using *ab Initio* Evolutionary Techniques: Principles and Applications. *The Journal of Chemical Physics*, **124**, Article ID: 244704. https://doi.org/10.1063/1.2210932

[3] Hautier, G., Jain, A., Mueller, T., *et al*. (2013) Designing Multielectron Lithium-Ion Phosphate Cathodes by Mixing Transition Metals. *Chemistry of Materials*, **25**, 2064-2074. https://doi.org/10.1021/cm400199j

[4] Rupp, M., Tkatchenko, A., Müller, K.R., *et al*. (2012) Reply to Comment on Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters*, **109**, Article ID: 059802. https://doi.org/10.1103/PhysRevLett.108.058301

[5] Snyder, J.C., Rupp, M., Hansen, K., *et al*. (2012) Finding Density Functionals with Machine Learning. *Physical Review Letters*, **108**, Article ID: 253002. https://doi.org/10.1103/PhysRevLett.108.253002

[6] Wolverton, C., Yan, X.Y., Vijayaraghavan, R., *et al*. (2002) Incorporating First-Principles Energetics in Computational Thermodynamics Approaches. *Acta Materialia*, **50**, 2187-2197. https://doi.org/10.1016/S1359-6454(01)00430-X

[7] Ceder, G. (1998) Predicting Properties from Scratch. *Science*, **280**, 1099-1100. https://doi.org/10.1126/science.280.5366.1099

[8] Moreels, P. and Perona, P. (2007) Evaluation of Features Detectors and Descriptors Based on 3D Objects. *International Journal of Computer Vision*, **73**, 263-284. https://doi.org/10.1007/s11263-006-9967-1

[9] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) Imagenet Classification with Deep Convolutional Neural Networks. In: *Advances in Neural Information Processing Systems*, Springer, New York, 1097-1105.

[10] Deng, L., Li, J., Huang, J.T., *et al*. (2013) Recent Advances in Deep Learning for Speech Research at Microsoft. 2013 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, 26-31 May 2013, 8604-8608. https://doi.org/10.1109/ICASSP.2013.6639345

[11] Braun, M.L., Buhmann, J.M. and MÃžller, K.R. (2008) On Relevant Dimensions in Kernel Feature Spaces. *Journal of Machine Learning Research*, **9**, 1875-1908.

[12] Geman, S., Bienenstock, E. and Doursat, R. (1992) Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, **4**, 1-58. https://doi.org/10.1162/neco.1992.4.1.1

[13] Vapnik, V. (2013) The Nature of Statistical Learning Theory. Springer Science & Business Media, New York.

[14] Cuingnet, R., Rosso, C., Chupin, M., *et al*. (2011) Spatial Regularization of SVM for the Detection of Diffusion Alterations Associated with Stroke Outcome. *Medical Image Analysis*, **15**, 729-737. https://doi.org/10.1016/j.media.2011.05.007

[15] Gaonkar, B. and Davatzikos, C. (2013) Analytic Estimation of Statistical Significance Maps for Support Vector Machine Based Multi-Variate Image Analysis and Classification. *NeuroImage*, **78**, 270-283. https://doi.org/10.1016/j.neuroimage.2013.03.066

[16] Boser, B.E., Guyon, I.M. and Vapnik, V.N. (1992) Proceedings of the Fifth Annual Workshop on Computational Learning Theory.

[17] Rosasco, L., Vito, E.D., Caponnetto, A., *et al*. (2004) Are Loss Functions All the Same? *Neural Computation*, **16**, 1063-1076. https://doi.org/10.1162/089976604773135104

[18] Hastie, T. and Tibshirani, R. (1998) Classification by Pairwise Coupling. In: *Advances in Neural Information Processing Systems*, Springer, New York, 507-513. https://doi.org/10.1214/aos/1028144844

[19] Hawkins, D.M. (2004) The Problem of Overfitting. *Journal of Chemical Information and Computer Sciences*, **44**, 1-12.

https://doi.org/10.1021/ci0342472

[20] Nair, V. and Hinton, G.E. (2010) Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the* 27*th International Conference on Machine Learning* (*ICML*-10), Haifa, Israel, 21-24 June 2010, 807-814.